Automatic Protein Function Prediction

Daniel Faria

Bioinformática

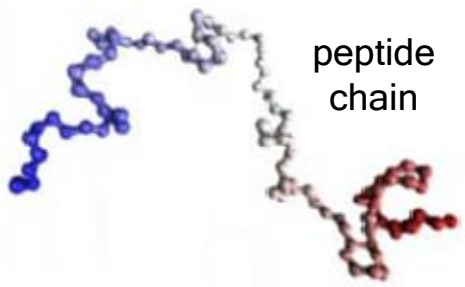
16 de Maio de 2011

The problem

DNA 

DNA Sequencing

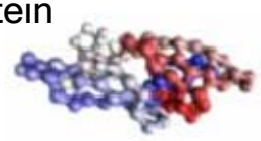
Transcription, Translation



peptide chain

Gene Finding

Folding

protein 

Crystallography

?

3D Structure → Molecular Function → Biological Process

Transcriptomics, Proteomics

State of the Art

Alignment-based inference (BLAST)

- Proteins with similar sequences likely evolved from a common ancestor (homologues), and likely have related functions
- Impossible to quantify; prone to propagate errors

Machine learning

- Use of robust machine learning algorithms to classify proteins
- Lack versatility; can’t deal directly with sequence data

Thesis

Stage 1

- Goal: develop a machine learning algorithm that can deal directly with sequences $\Rightarrow$ Peptide Programs: modelling sequences as small computer programs
- Task: binary classification of 18 enzyme families

	BLAST	**Property SVMs**	**Peptide Programs**
Precision	**99%**	96%	98%
Recall	**98%**	97%	94%
Matthews CC	**0,98**	0,96	0,95

Thesis

Stage 2

- Goal: find alternative sequence parametrizations that encode as much information as possible
- Task: binary classification of 18 enzyme families

	BLAST	Property SVMs	6 Part SVMs	Dipeptide SVMs	Tripeptide SVMs
Precision	99%	96%	99%	99%	**100%**
Recall	**98%**	97%	96%	96%	96%
Matthews CC	**0,98**	0,96	0,97	0,97	**0,98**

Thesis

Stage 3

- Goal: explore descriminating power of tripeptides $\Rightarrow$ TripSim: tripeptide-based sequence similarity
- Task: binary classification of 18 enzyme families

	BLAST	TripSim
Precision	**99%**	**99%**
Recall	**98%**	**98%**
Matthews CC	**0,98**	**0,98**

Thesis

Stage 4

- Goal & Task: predict detailed molecular function annotations of proteins
- Approaches: BLAST or TripSim search followed with prediction by nearest neighbours (NNb) or Tripeptide-based SVMs

	BLAST		TripSim	
	NNb	**Trip SVMs**	**NNb**	**Trip SVMs**
Semantic Similarity	47%	51%	49%	51%

Stage 5

Write Thesis!

Automatic Protein Function Prediction

Bioinformática

16 de Maio de 2011

Keeping up with the research

Automated extension of biomedical ontologies

Catia Pesquita
LaSIGE
Department of Informatics, FCUL

May 16, 2011

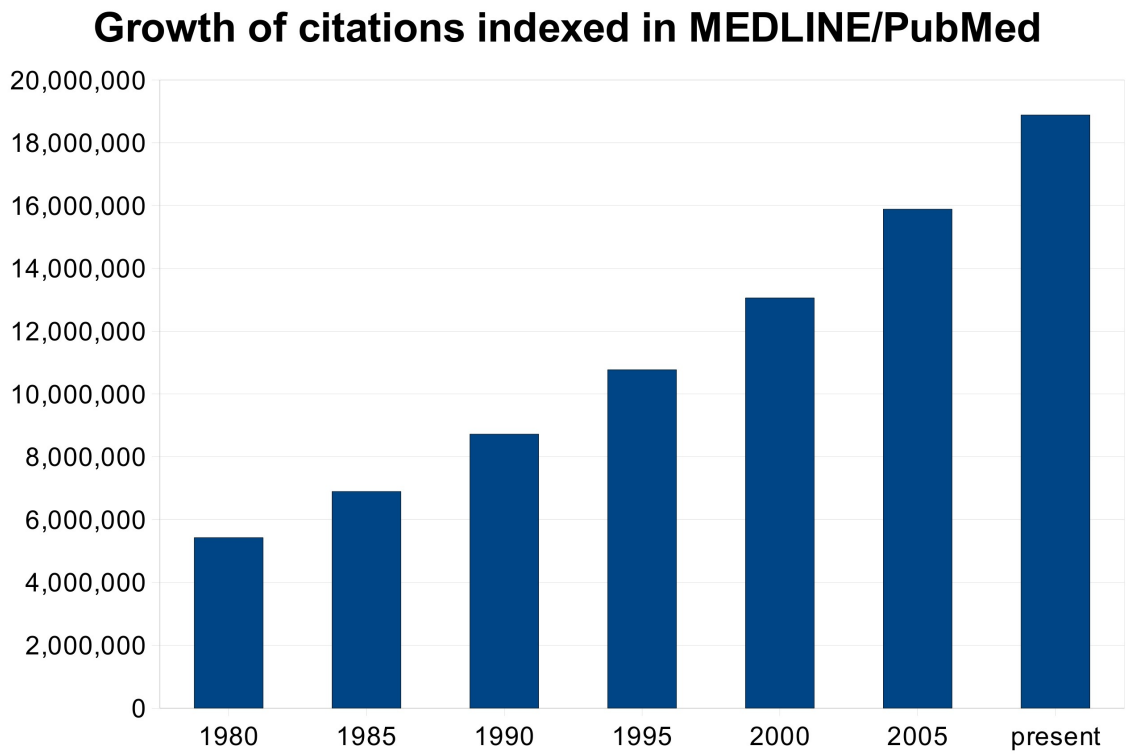
Developing a biomedical ontology

Q: How many people does it take to finish a biomedical ontology?

Developing a biomedical ontology

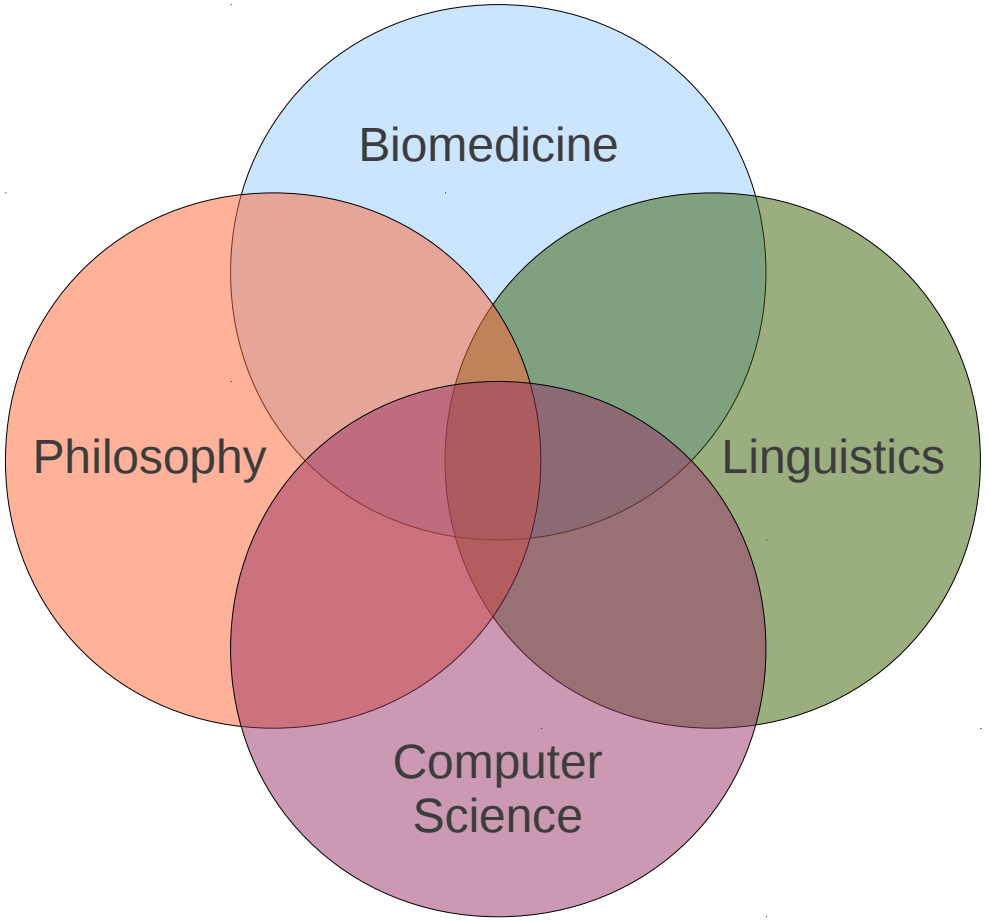
Q: How many people does it take to finish a biomedical ontology?

A: A biomedical ontology is never complete.

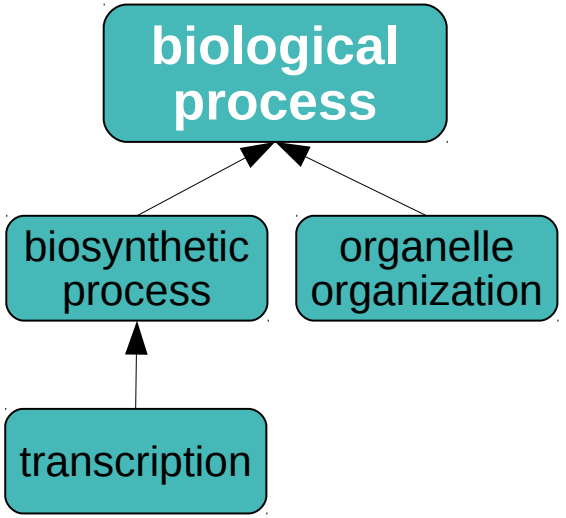


Developing a biomedical ontology

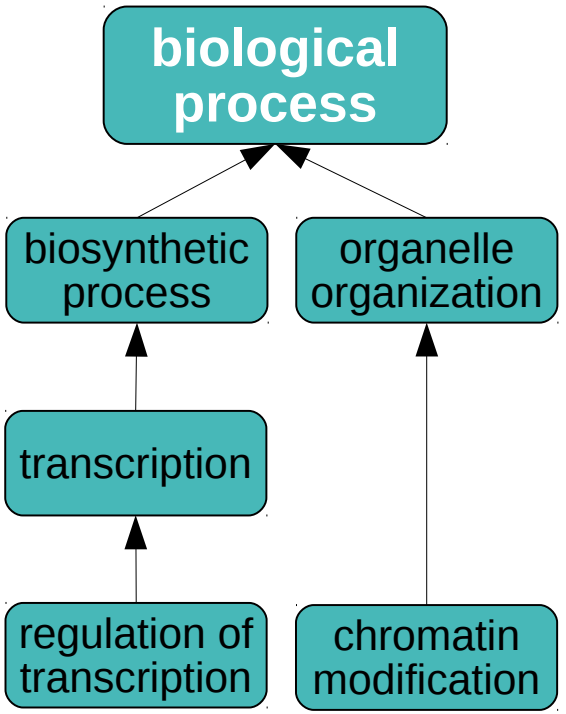
- **Keep up with the research in the domain**
- Art -> Science: quality metrics, scientific techniques
- Experts in several areas



Ontology Extension



Ontology Extension



- What? Add new classes and relations
- Why? New knowledge, new point of view, etc...
- How? User requests, **read a lot of papers**, ...

Automated Ontology Extension

For this we need KNOWLEDGE, which can be found in...

Automated Ontology Extension

For this we need KNOWLEDGE, which can be found in...

TEXT
- High coverage
- Unstructured

Automated Ontology Extension

For this we need KNOWLEDGE, which can be found in...

TEXT
- High coverage
- Unstructured

OTHER ONTOLOGIES
- Lower coverage
- Structured

Automated Ontology Extension

For this we need KNOWLEDGE, which can be found in...

TEXT
- High coverage
- Unstructured

Text Mining

OTHER ONTOLOGIES
- Lower coverage
- Structured

Ontology Matching

Auxesia

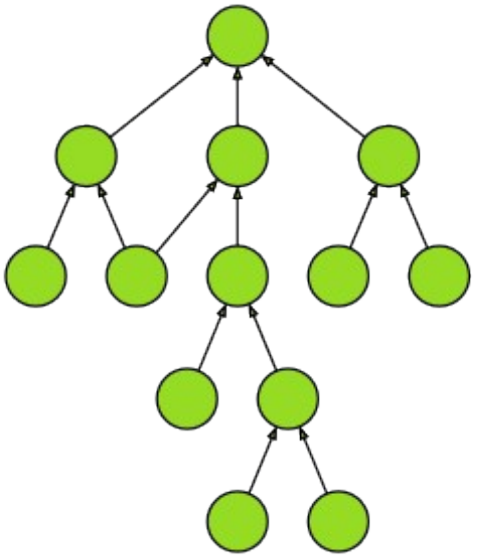
- Greek goddess of spring growth

- Methodology for semi-automated biomedical ontology extension

- Two main challenges:
  1. Bio-ontologies are LARGE
  2. Bio terminology is COMPLEX

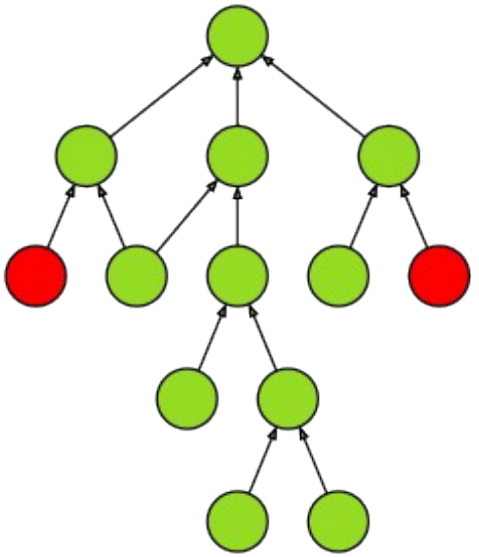
- Two novel ideas:
  1. Predict the areas of ontology that need to be extended
  2. Explore biomedical annotation corpora in Text Mining and Ontology Matching → semantic similarity

Auxesia



Auxesia

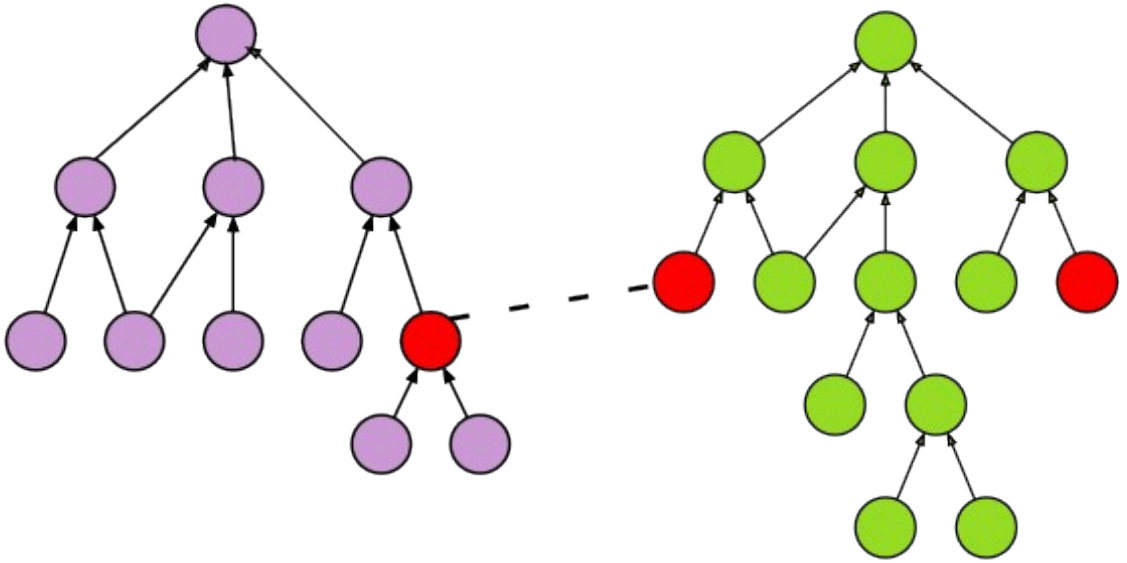
1. Predict ontology areas for extension



Auxesia

1. Predict ontology areas for extension

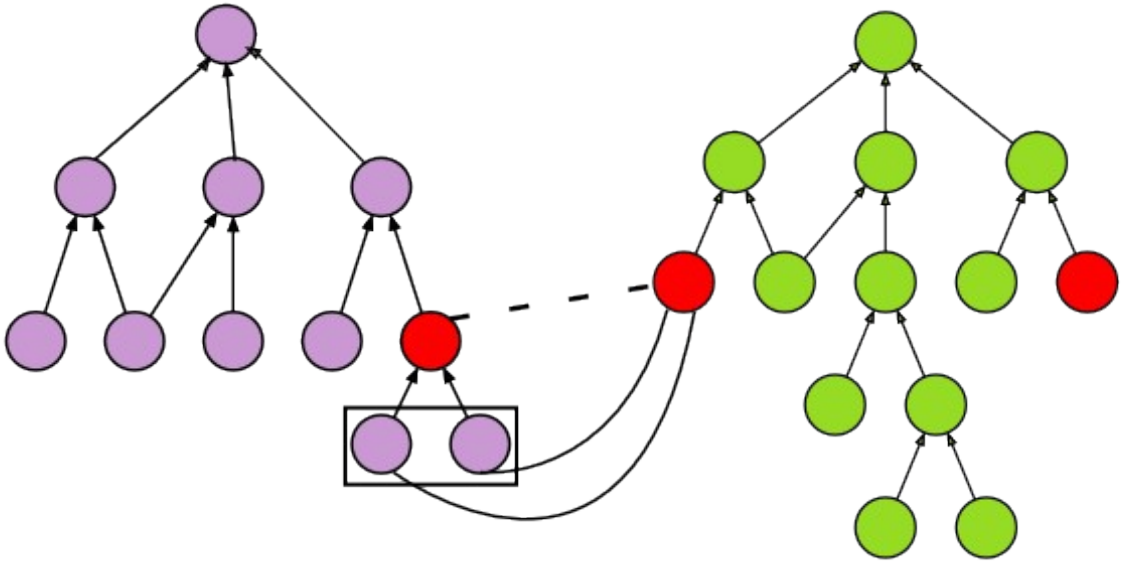
2.1 Match to relevant ontologies



Auxesia

1. Predict ontology areas for extension

2.1 Match to relevant ontologies



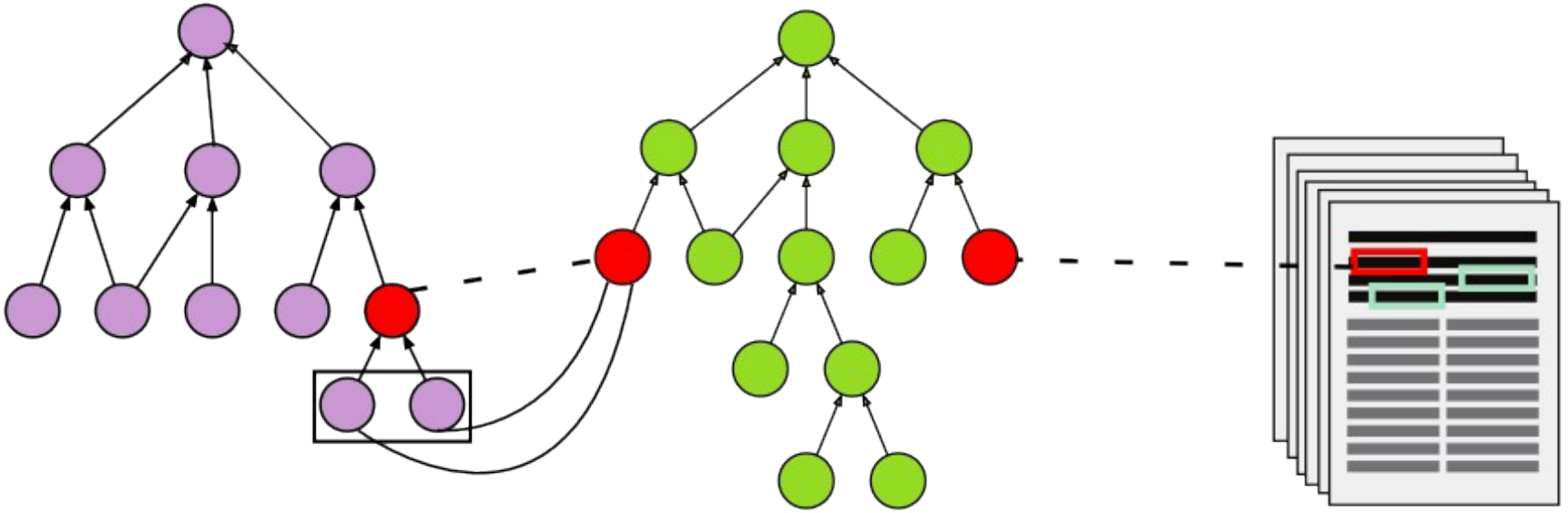
2.2 Propose new candidate classes

Auxesia

1. Predict ontology areas for extension

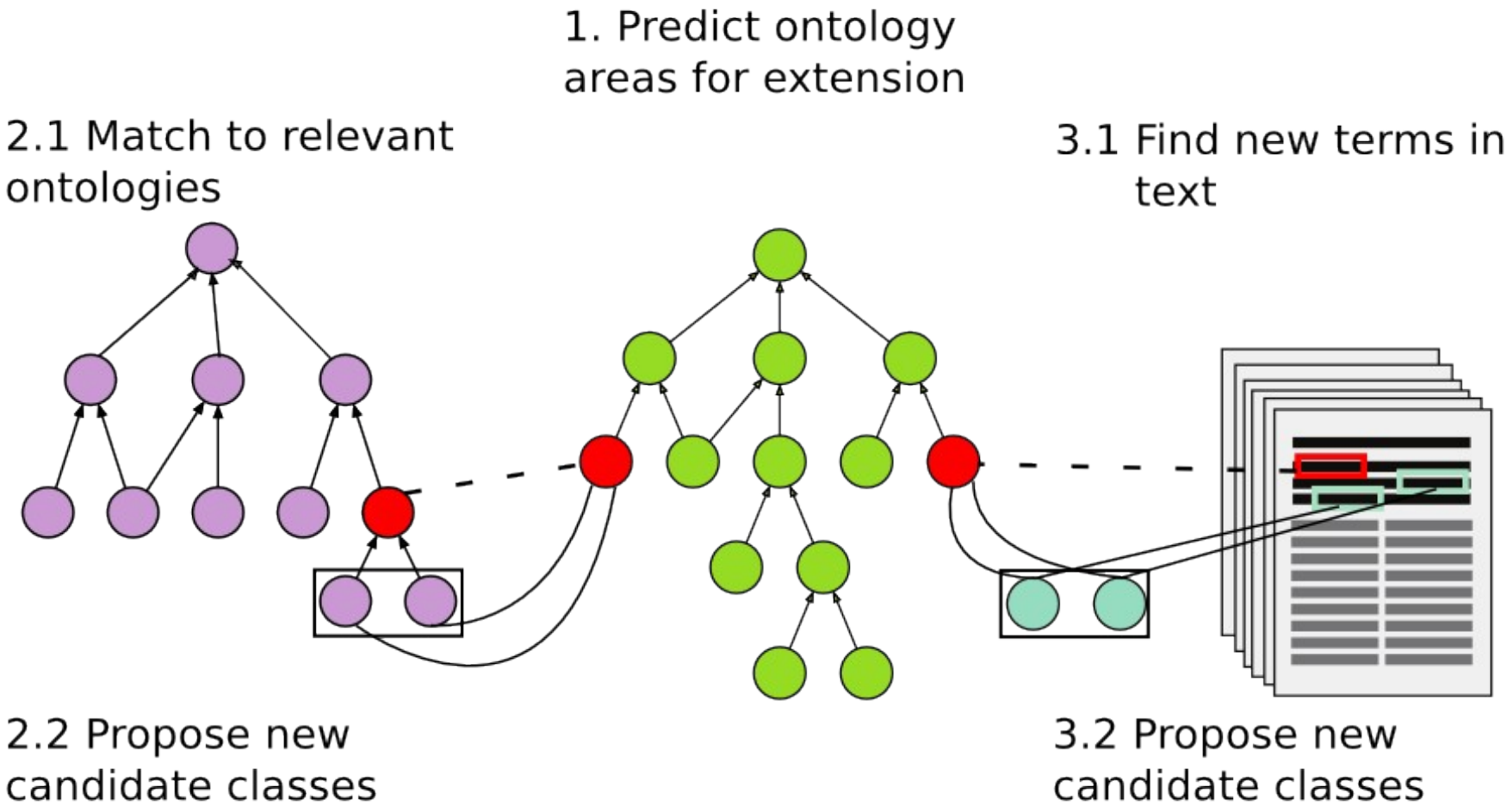
2.1 Match to relevant ontologies

3.1 Find new terms in text



2.2 Propose new candidate classes

Auxesia


1. Predict ontology areas for extension
2.1 Match to relevant ontologies
3.1 Find new terms in text
2.2 Propose new candidate classes
3.2 Propose new candidate classes

The story so far...

- **Prediction of ontology extension – Novel area**:
  - Gene Ontology
  - Machine Learning based on ontology and annotation corpora statistics
  - 70% success

- **Ontology Matching**
  - Anatomical Ontologies
  - Lexical and Structural methods
  - 2nd place in precision in OAEI (within 0.001 of 1st)

- **Text Mining**
  - Still working on it...

Would you like to know more?

- http://xldb.di.fc.ul.pt/wiki/AEB
- C6 3.30

Identifying Chemical Entities

Tiago Grego

Universidade de Lisboa

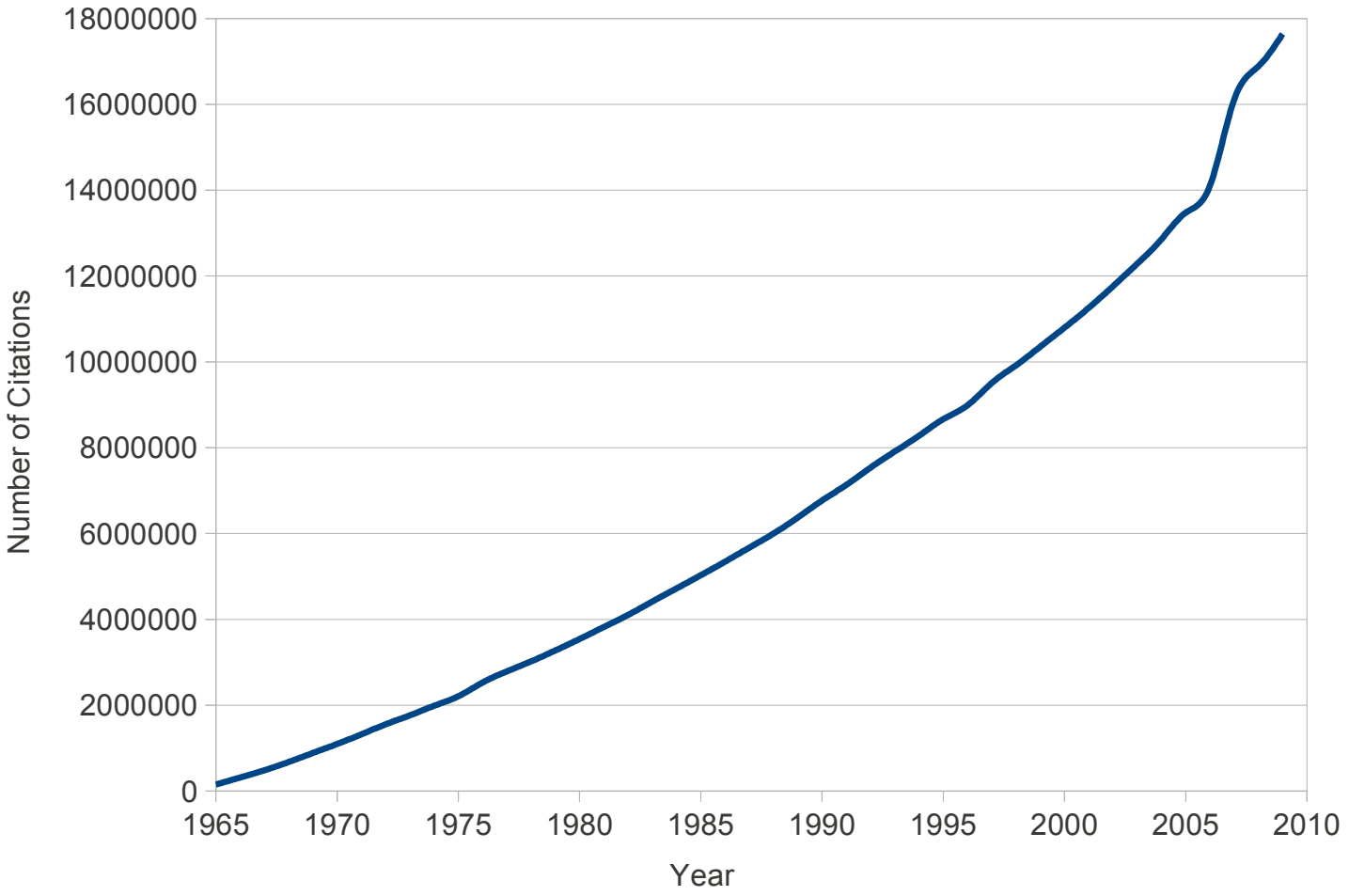
16th May 2011

Motivation

- **Increasing number of publications**
  - Journal Articles
  - Patents
  - Reports
  - Books
  - …
- **High Throughput methods**
  - Automated DNA sequencing
  - DNA expression arrays
  - Mass spectrometry
  - High Throughput screening
  - …

Literature

- **Growth of MEDLINE**



http://www.nlm.nih.gov/bsd/index_stats_comp.html

Motivation

- Far too much literature to review or keep up-to-date manually.
- Tools are needed to organize, curate and retrieve chemical information

Text Mining

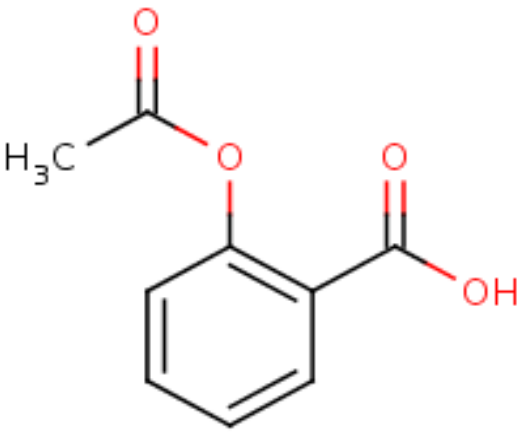
- First step is Named Entity Recognition

Challenges

- Infinite number of possible chemical compounds

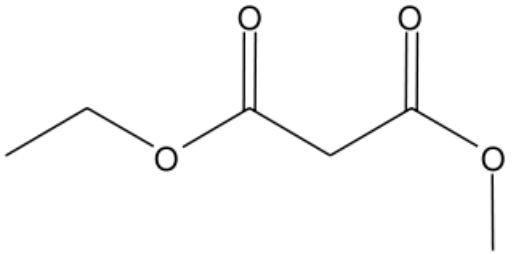
Challenges

- Infinite number of possible chemical compounds
- No standard for chemical representation
  - Trivial names: Aspirin, acetylsalicylate, acetylsalicylic acid
  - Systematic names: 2-acetyloxybenzoic acid, 2-(acetyloxy)benzoic acid
  - IUPAC name: pyrido[1",2":1’,2’]imidazo[4’,5’:5,6]pyrazino[2,3-b]phenazine
  - SMILES: OC(=O)C1=C(C=CC=C1)OC(=O)C
  - InChI: 1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)/f/h11H
  - Empirical formula: CH3OOC6H4COOH
  - Mass formula: C9H8O4
  - CAS registry number: 50-78-2
  - Abreviation: ASA

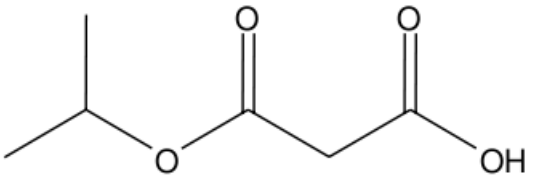


Challenges

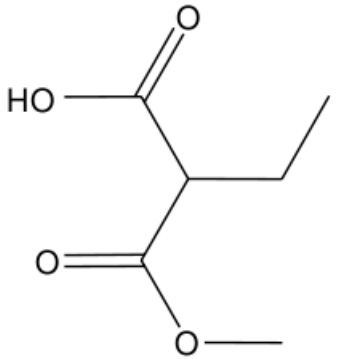
- Infinite number of possible chemical compounds
- No standard for chemical representation
- High ambiguity and polysemy



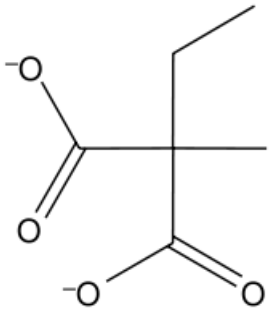
Methyl ethyl malonate



Methylethyl malonate



Methyl ethylmalonate



Methylethylmalonate

Motivation

- Drug Discovery:
  - 90% of drug targets are derived from the literature
    - Which proteins interact with my target protein?
    - Who is patenting on what targets, compounds and diseases?
    - Will my lead compound fail due to toxicity?
    - ...

Motivation

- Drug Discovery:
  - 90% of drug targets are derived from the literature
    - Which proteins interact with my target protein?
    - Who is patenting on what targets, compounds and diseases?
    - Will my lead compound fail due to toxicity?
    - ...
  - 50% of all potentially therapeutic compounds fail due to toxicity

Motivation

- Drug Discovery:
  - 90% of drug targets are derived from the literature
    - Which proteins interact with my target protein?
    - Who is patenting on what targets, compounds and diseases?
    - Will my lead compound fail due to toxicity?
    - ...
  - 50% of all potentially therapeutic compounds fail due to toxicity

50% of them had some indication in the literature already

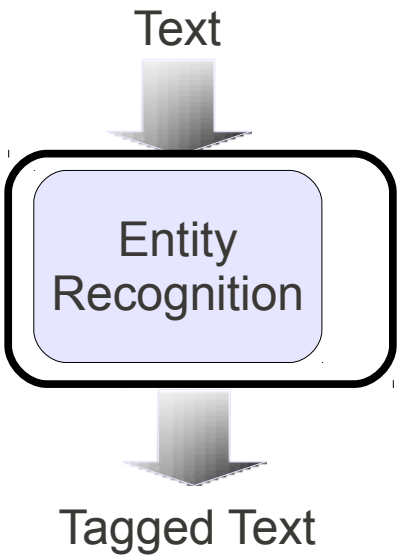
Fickett, J. and Hayes, W. (2004) Text Mining for Drug Discovery. European Pharmaceutical Contractor.

Available Systems

Name	Resources used	Approaches
Whatizit	ChEBI	NLP
OSCAR3	Internal dictionary based in ChEBI	NLP, low rule-based
Bioalma AKS	PubChem	NLP
SCAI	Several dictionaries	NLP for trivial names, ML for systematic names
BioTeKS	Several dictionaries and ontologies	NLP, ML, Rule-based
I2E	Domain specic ontologies	NLP
Luxid	Several dictionaries	NLP and Rule-based

- Most systems are Dictionary-based

Overview



1) Entity Recognition Module – Machine Learning

Entity Recognition Module
Corpus (Gold Standard)

- A set of 40 patent documents, selected as a representative set of the chemical patent documents, was manually curated by a joint team of EPO and ChEBI curators
  - 18061 annotations (10170 mapped to ChEBI)

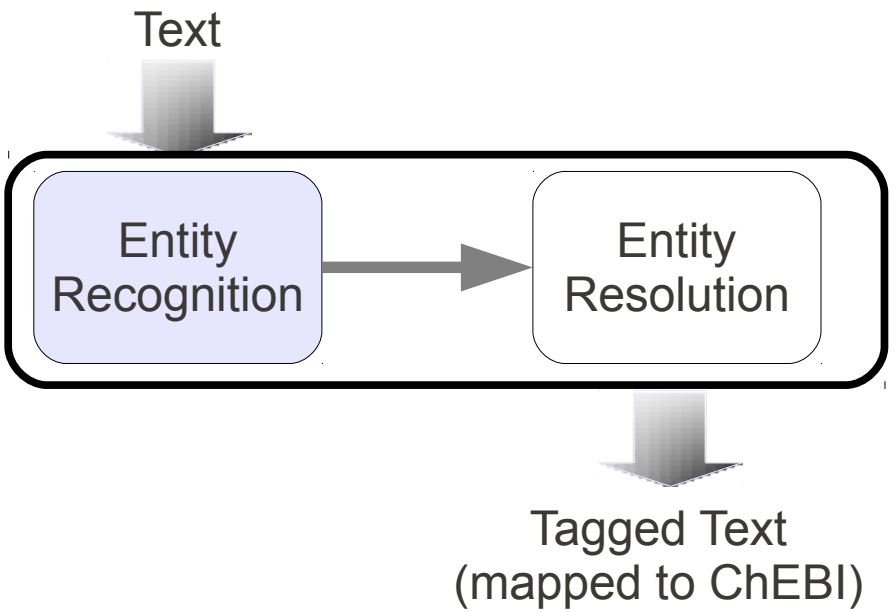
- Using this corpus, a Case-based chemical entity recognition method was developed and compared to Dictionary-based methods
  - WhatizitChebiDict

Entity Recognition Module
Case-based approach

- Conditional Random Fields
    - Probabilistic framework for labeling and segmenting structured data, such as sequences

- MALLET (MAchine Learning for LanguagE Toolkit) implementation (http://mallet.cs.umass.edu)

Overview



1) Entity Recognition Module
2) Entity Resolution Module

Entity Resolution Module

- It's important to find the entities...
  … but also to know which entities you found.

- Implicit in Dictionary-based entity recognition
  but not with the developed Case-based approach.

- Search entity in database, more weight

NER and Resolution

Results

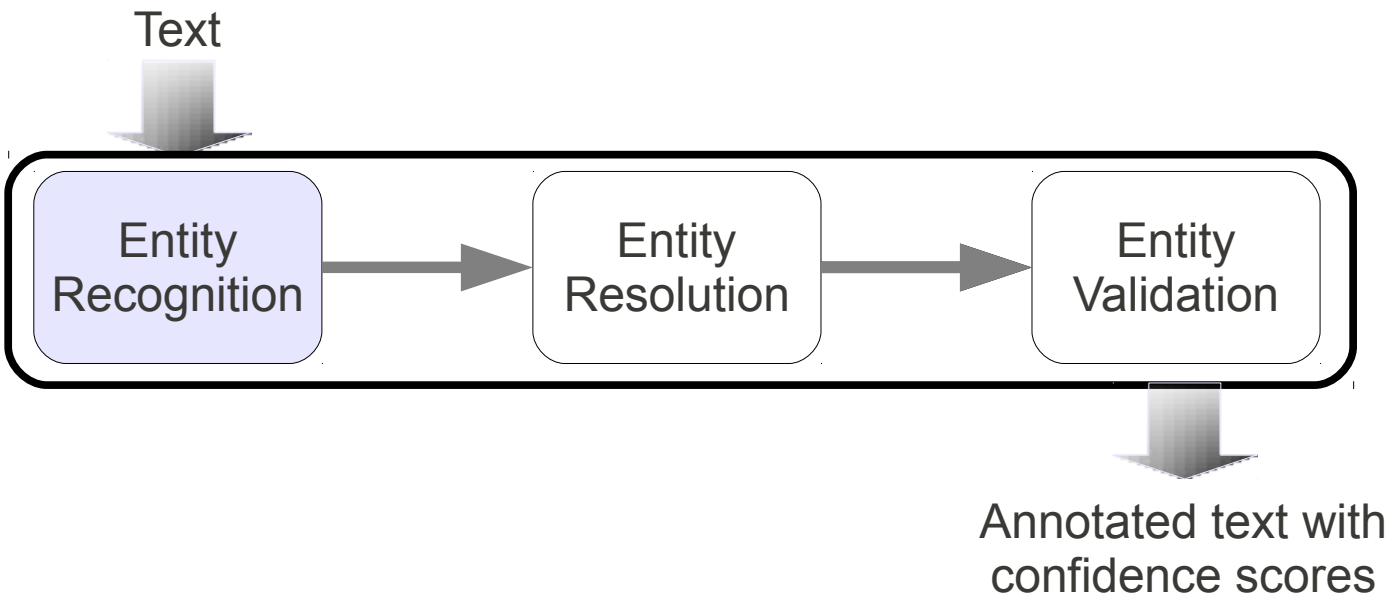
- NER

		Manual	Automatic	Precision	Recall	F-measure
Exact matching	Dictionary lookup	18061	18683	29.78	30.80	30.28
	Machine learning	18061	13832	61.01	46.72	**52.92**
Partial matching	Dictionary lookup	18061	18683	54.88	56.77	55.81
	Machine learning	18061	13832	86.87	66.53	**75.35**

- NER + Resolution

		Manual	Automatic	Precision	Recall	F-measure
Exact matching	Dictionary lookup	10170	18683	23.11	42.46	29.93
	Machine learning	10170	10681	41.58	43.67	**42.60**
Partial matching	Dictionary lookup	10170	18683	27.23	50.02	35.26
	Machine learning	10170	10681	48.10	50.52	**49.28**

Overview



1) Entity Recognition Module
2) Entity Resolution Module
3) Entity Validation Module

Entity Validation
Introduction

- The scope of a scientific publication is typically narrow
  - Specific protein, specific metabolic pathway, specific disease, etc

- Thus it is expected for the entities present in a document to be related

- Relationships found between recognized entities provide evidence of correct entity recognition

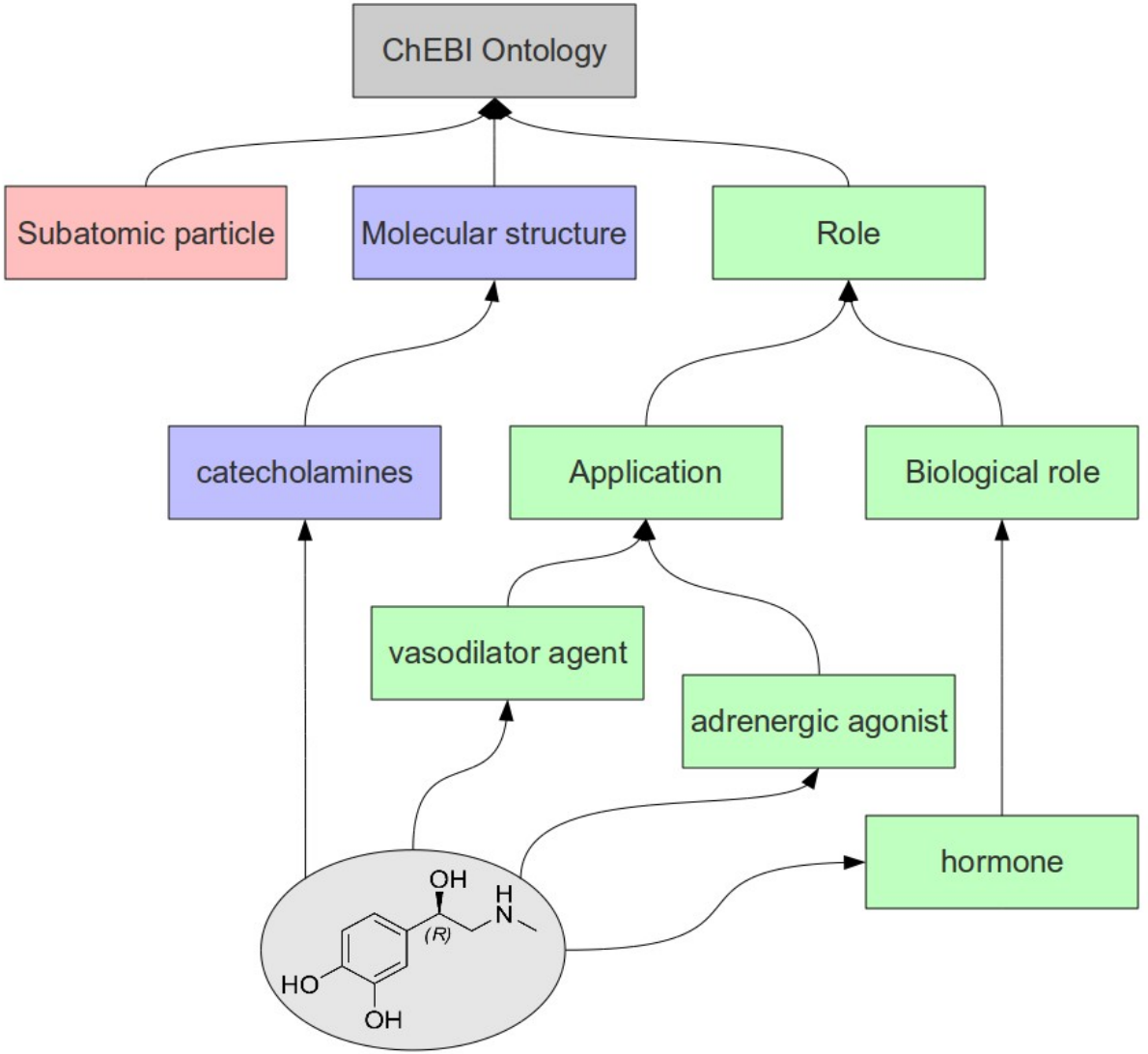
- No relationships found for an entity are evidence of a recognition error

Entity Validation
Approach

- Comparison of Chemical Compounds
  - Structural Similarity
  - Semantic Similarity

- ChEBI Ontology
  - Structural Similarity
  - Functional Similarity

Entity Validation
Approach



Example

This is a prototype of ICE.

This is a sample text containing chemical compounds such as water and oxygen.

Other famous chemical entities are ethanol, also called ethyl alcohol.

Some chunks of text such as metanyl alcahol sound like a chemical entity but aren't.

Send Reset

Example

Result

This is a sample text containing chemical compounds such as water and oxygen.

Other chemical entities can include ethanol, also called ethyl alcohol.

Some chunks of text such as metanyl alcahol sound like a chemical entity but aren't.

Back

EMBL-EBI Enter Text Here Find Help | Feedback

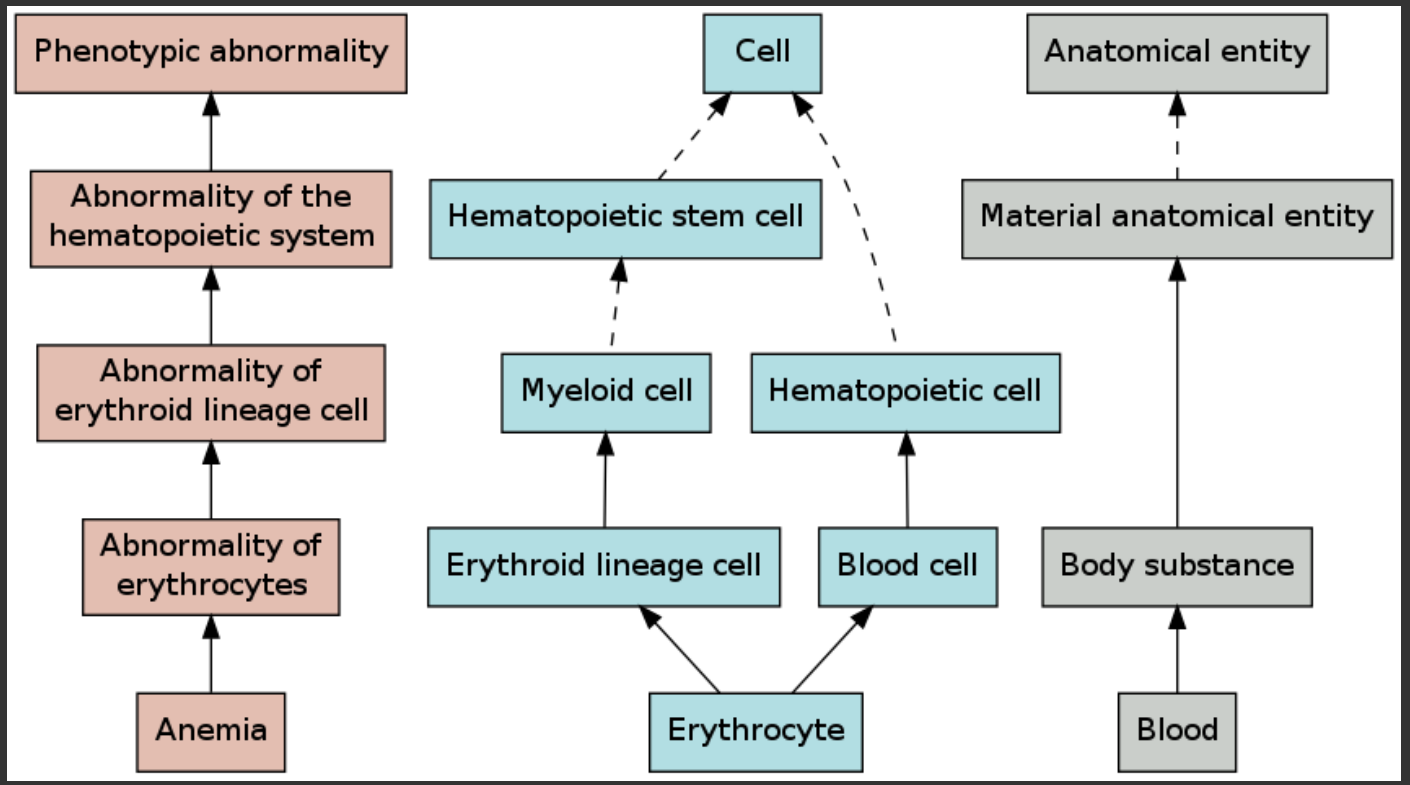
Databases | Tools | Research | Training | Industry | About Us | Help | Site Index

Resulting tagged text

This is a sample text containing chemical compounds such as water and oxygen. Other chemical entities can include ethanol, also called ethyl alcohol. Some chunks of text such as metanyl alcahol sound like a chemical entity but aren't .

- calcium(0) (CHEBI:29320) contains as synonym "Ca*n*"
- New chemicals cannot be identified with dictionary-based methods

Working with multiple biomedical ontologies



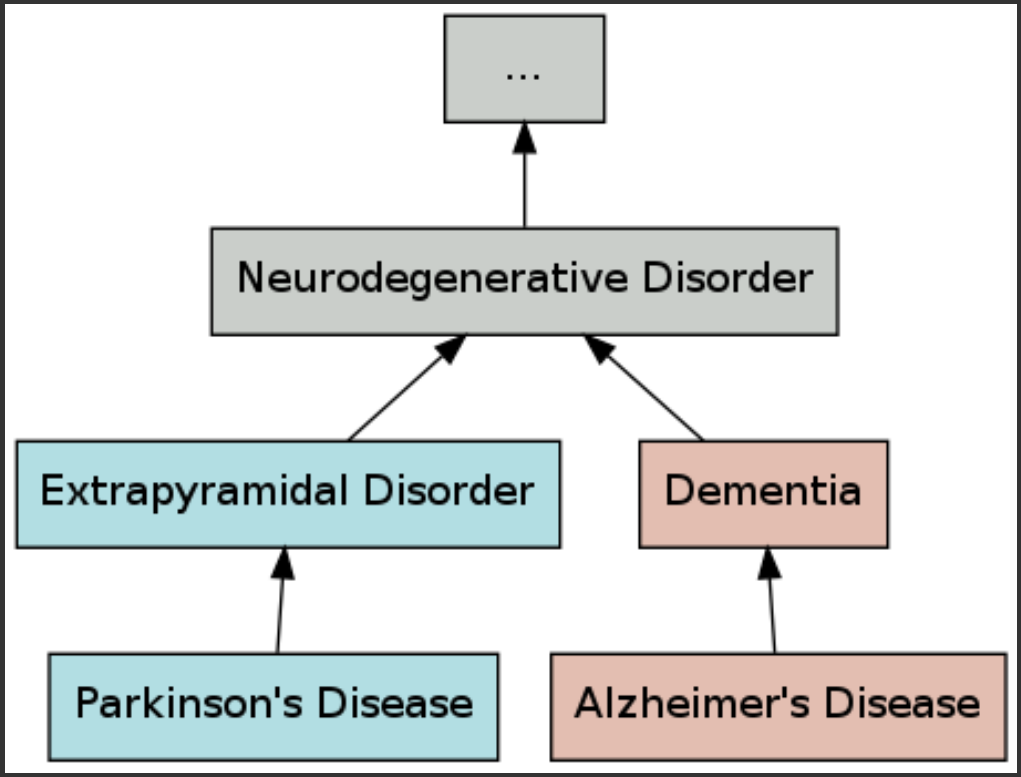
João D. Ferreira
PhD Student

May 16th, 2011

Ontology usage

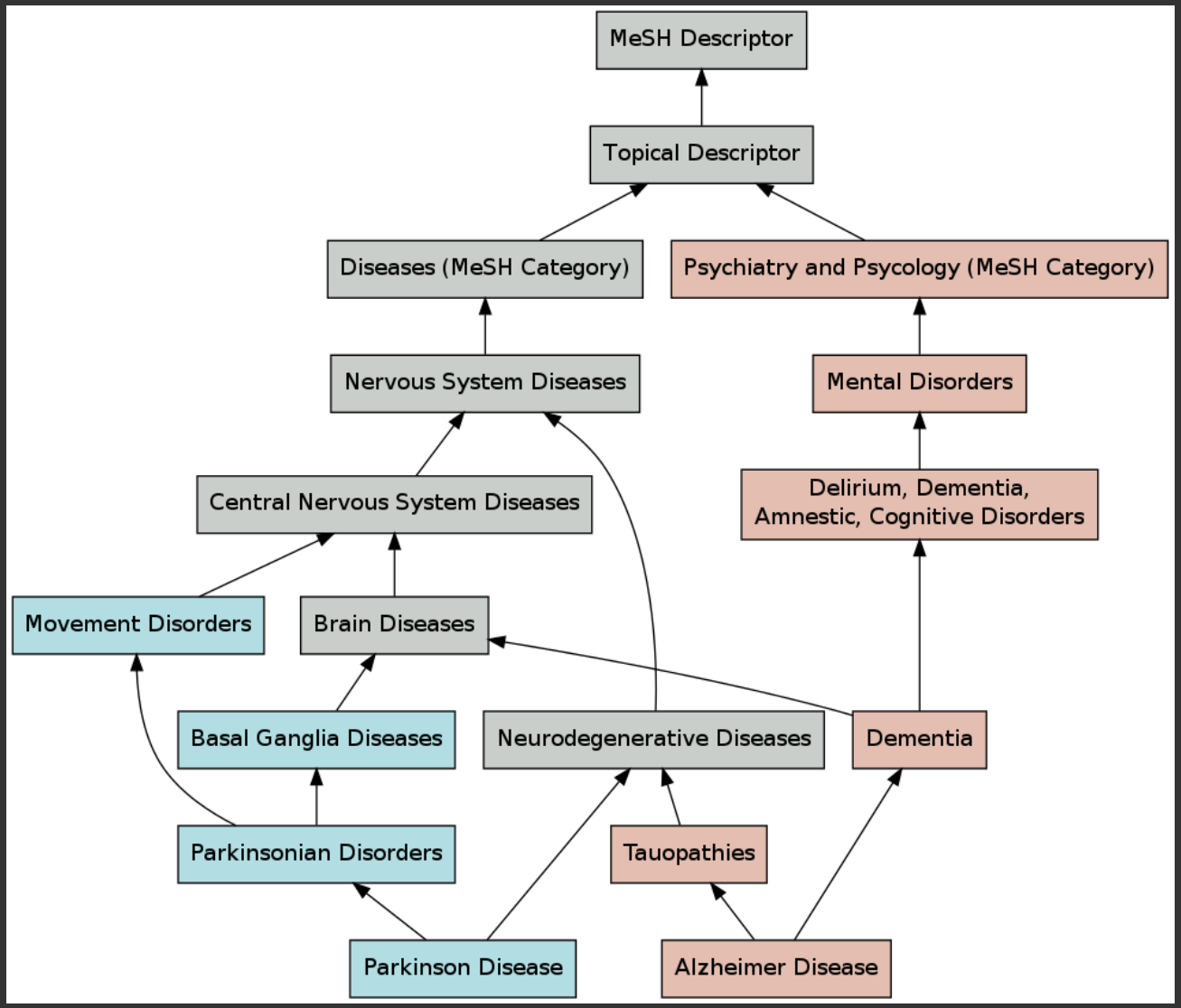
Information Catalog

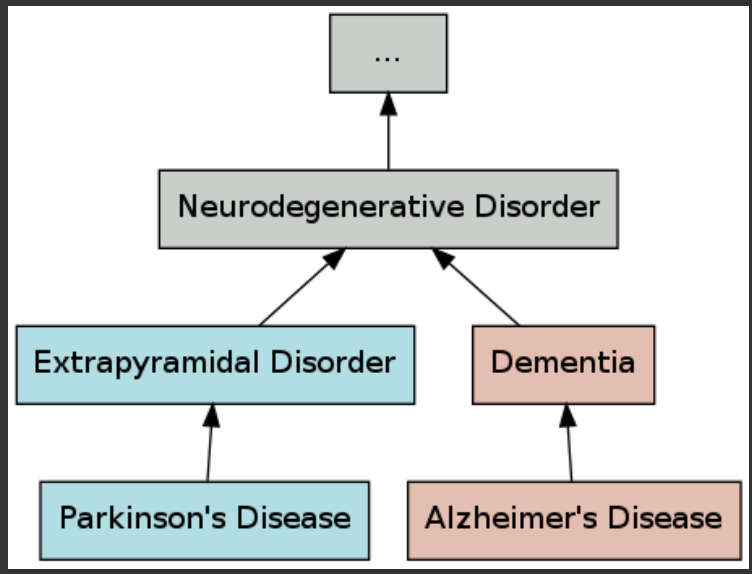
- Information becomes *computable*
  - Automatic reasoning
  - Semantic similarity
  - Pattern recognition
  - Text mining
  - Annotation of resources
    - *Papers*
    - *Web pages*
    - *Data sets*
    - *...*



Why multiple ontologies?

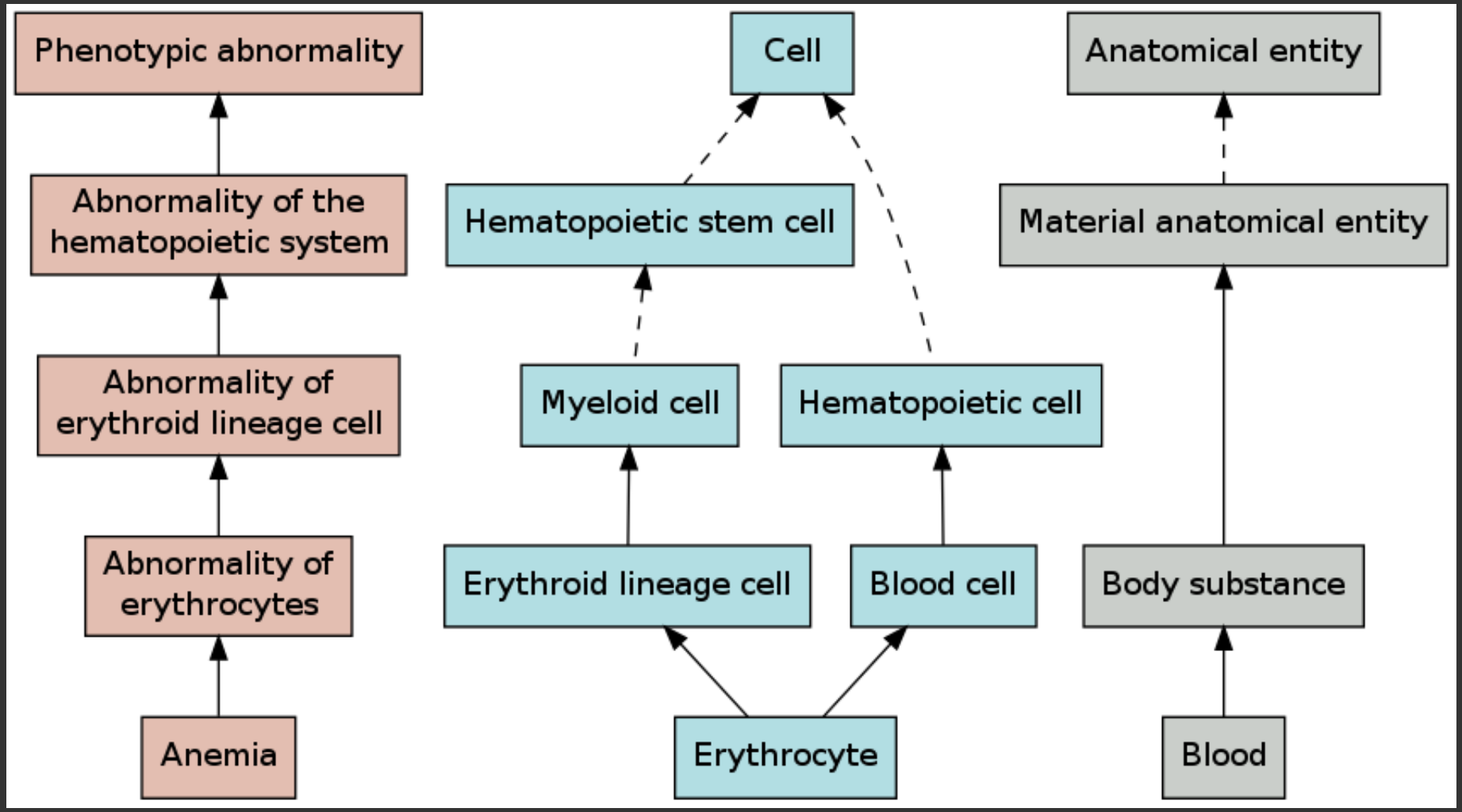
Different points of view on a subject





Why multiple ontologies?

Different domains of knowledge



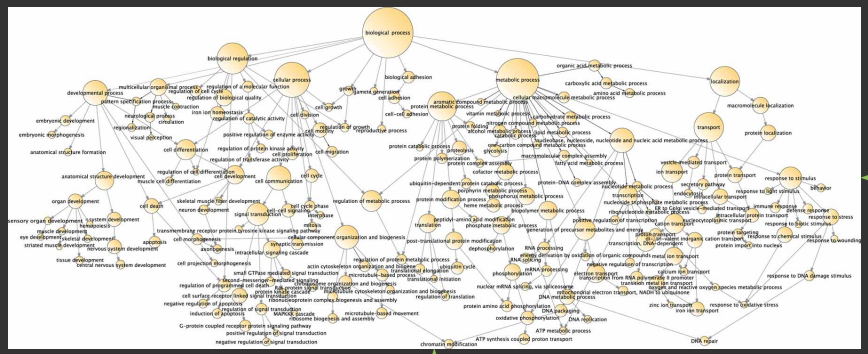
Integration

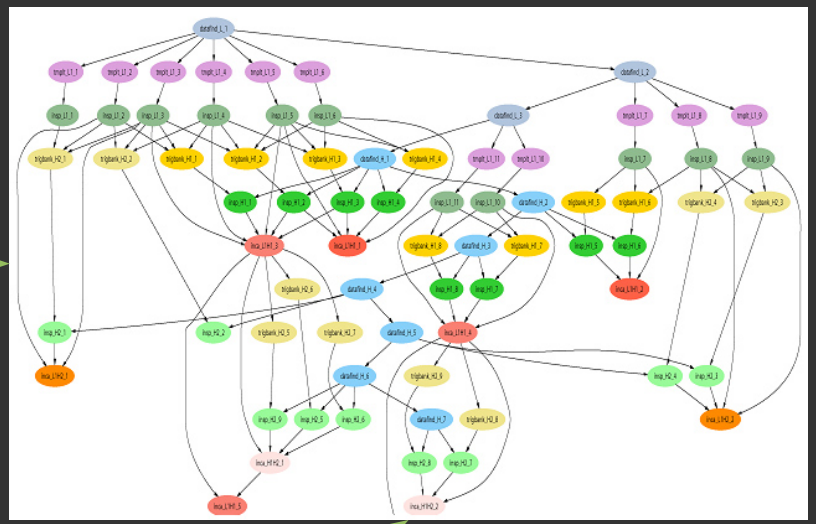
External references

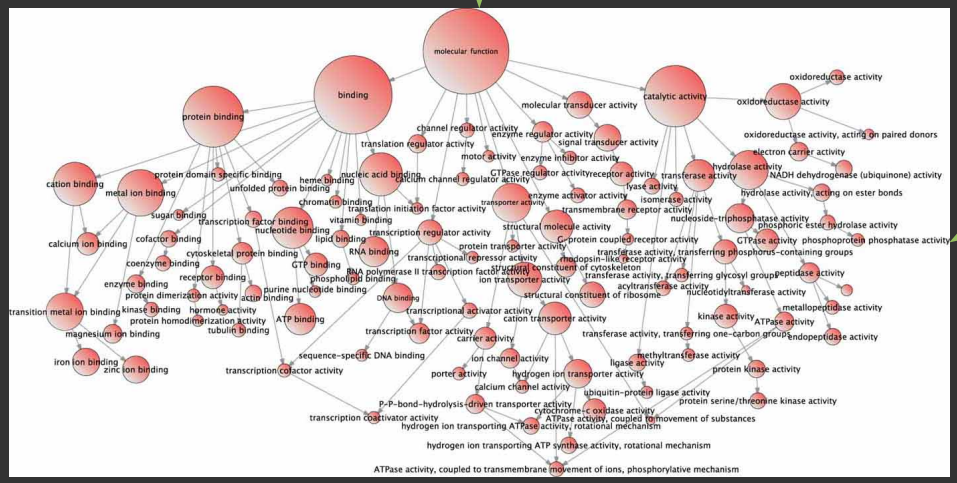
- Anemia (HPO) ⟶ Erythrocyte (CL) ⟶ Blood (FMA)
- Alzheimer's (HPO) ⟶ Brain (FMA)
- Parkinson's (HPO) ⟶ dopamine (PR), α-synuclein (PR)
- Cystic Fibrosis (HPO) ⟶ Lung (FMA), CFTR (PR)
- ...

Integration

External references

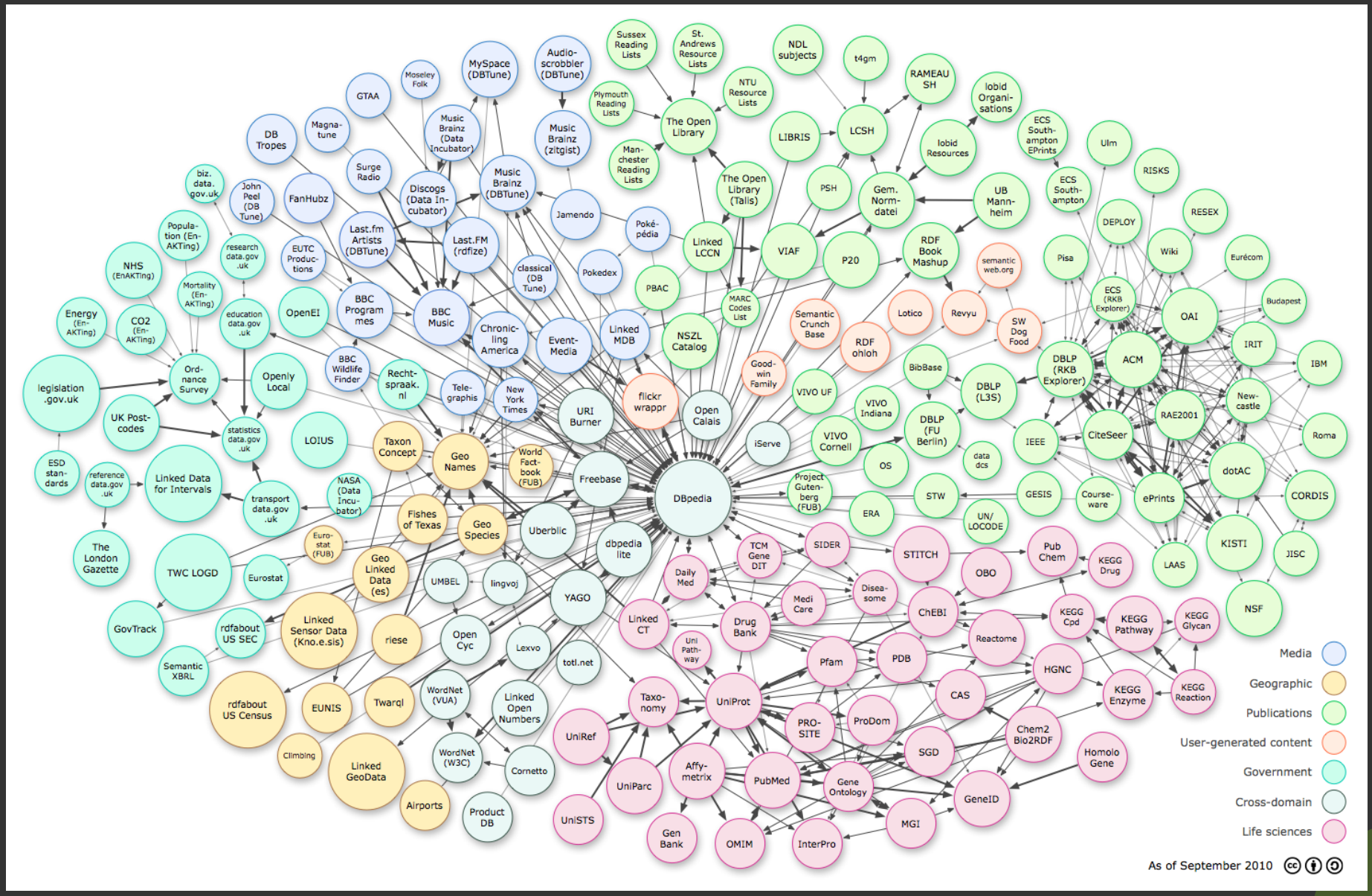






Integration

External references



Leading into ...

- More detail in the representation of knowledge
- Better and wider description of models
- Increased performance of automatic methodologies:
  - Search
  - Semantic similarity / relatedness
  - Automatic reasoning
  - ...

Applications

Epidemic Marketplace

- Repository for epidemiological resources

Applications

Epidemic Marketplace

- Annotation of resources with concepts from multiple ontologies:
  - Diseases
  - Phenotypes
  - Symptoms
  - Methods of transmission
  - Anatomical entities
  - Pathogens and vectors
  - Causes
  - Treatments
  - Geographical locations
  - …

Applications

Epidemic Marketplace

Example:

This report mentions the consequences of the outbreak of the 2009 Influenza A (H1N1) pandemic in Europe and North America.
This kind of influenza is transmitted through droplets from coughing and sneezing, and manifests though fever, gastrointestinal symptoms and possible pneumonia.

Applications

Epidemic Marketplace

Example:

This report mentions the consequences of the outbreak of the 2009 Influenza A (H1N1) pandemic in Europe and North America.
This kind of influenza is transmitted through droplets from coughing and sneezing, and manifests though fever, gastrointestinal symptoms and possible pneumonia.

Applications

Epidemic Marketplace

- *Annotation* of resources with concepts mentioned in the text
- An advanced *search tool* is able to efficiently search all resources based on the annotations
- *Semantic relatedness of resources* used to retrieved resources which the user may also be interested in

Applications

Epidemic Marketplace

Blood diseases

Search

Applications

Epidemic Marketplace

Blood diseases | Search

- Resources mentioning:
  - *Blood*
  - *Erythrocytes*
  - *Anemia*
  - *Hemoglobin*
  - …

Collecting, Retrieving, and Estimating Reliable Thermochemical Data

Ana Isabel Teixeira

[ateixeira@lasige.fc.ul.pt]

16/05/2011

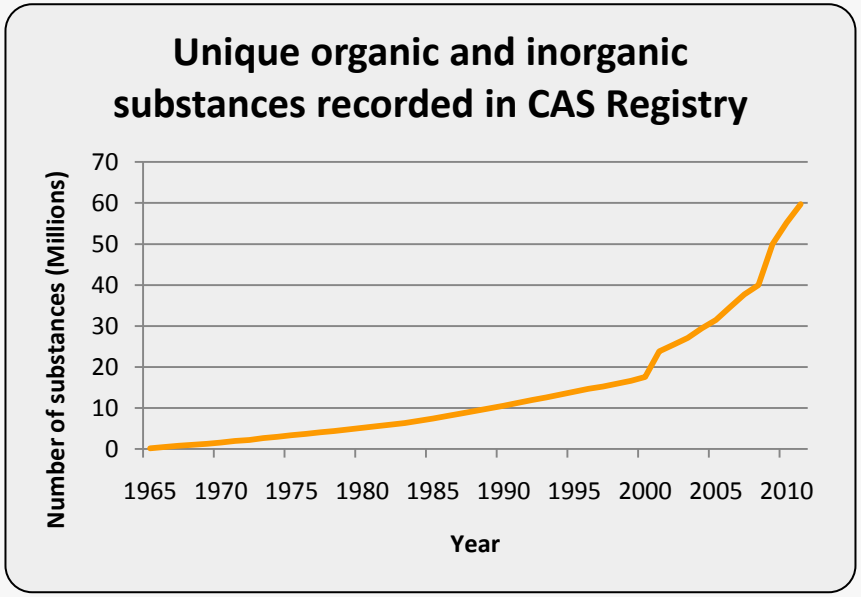
Outline

- Background
- System Architecture
- ThermInfo main features:
  - Searching for compounds
  - Predicting thermochemical properties
- Prediction of thermochemical properties of chemical compounds using machine learning techniques

Background

Data Avalanche in Chemistry

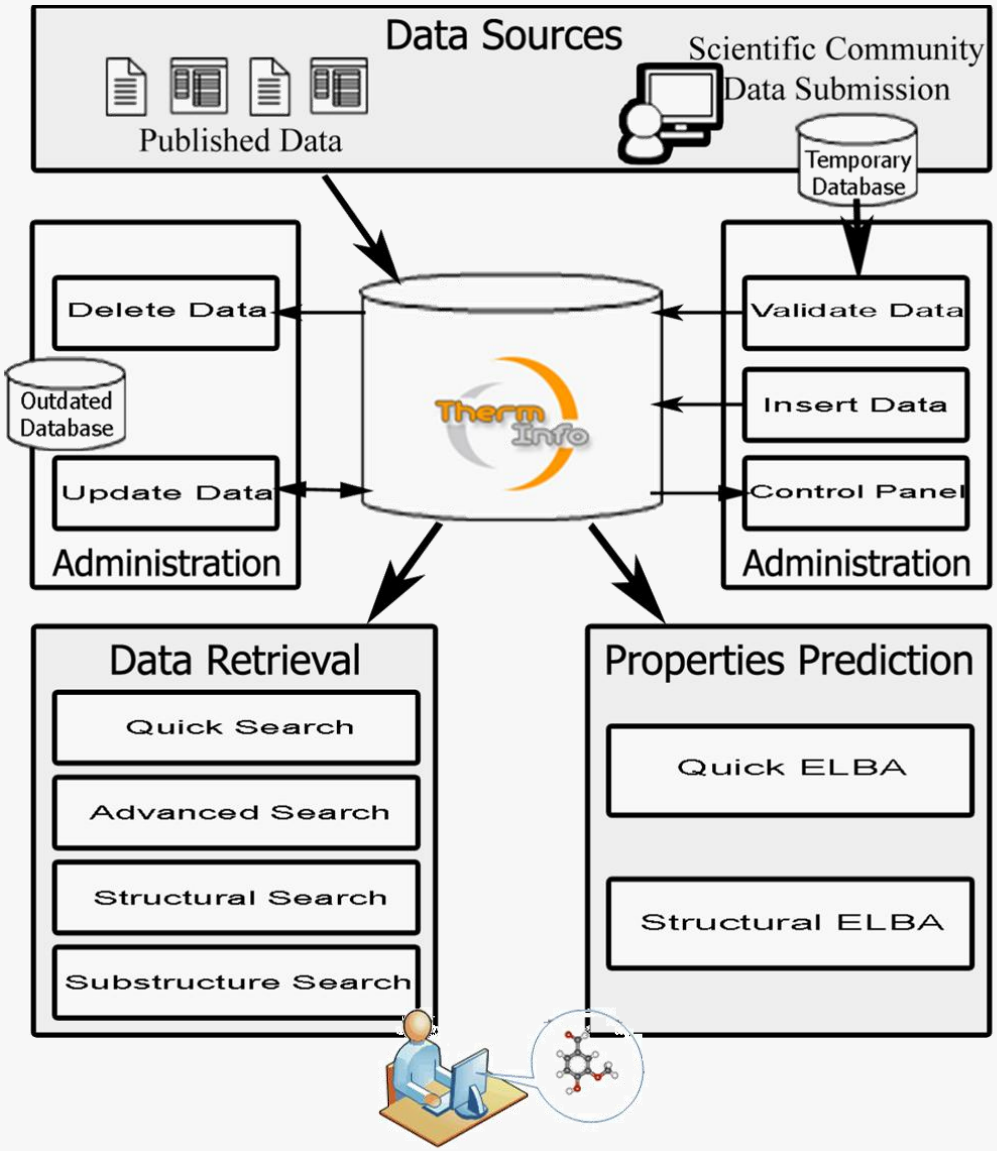
- + 59 million chemical compounds



Problems:

1. **Find Chemical Data**
   - Books and papers: complex and time-demanding;
   - Commercial databases: expensive;
   - Public databases: lack of data and/or critical evaluation.
2. **Availability and experimental determination of thermochemical data**

System Architecture



ThermInfo features: Searching for compounds

Search [Help]

Name | Search

Please do no

- Name
- Molecular Formula
- Molecular ID
- CAS RN
- SMILES

Security code:

[Type only **numerical characters**. Ignore letters and special characters.]

This Search Method was already used 101 times

Structural Search [Help]

Arquivo Editar Ver Atom Bond Tools Templates Ajuda

C H O N P S F Cl Br I +1 -1

Load SMILES | Clear | Identical Structures | Search

- Identical Structures
- Similar Structures, >=95%
- Similar Structures, >=90%
- Similar Structures, >=80%
- Similar Structures, >=70%

This Search Method was already

Advanced Search [Help]

Compound Name: | Physical State: All

Molecular Formula: | Molecular Weight: =

[Please type Molecular Formula with atoms in **CHXNOS** (X = halogen) order]

Wildcard: **?** represents one character

SMILES: | Identical Structures

Class: All

Sub-Class: All

Family: All

Characteristic:

CH Groups	Alkane [R-H]	Alkene [R2C=CR2]	Alkyne [R-C≡C-R]	Arene [Ar-H]
CHO Groups	Alcohol [R-OH]	Ether [R-O-R]	Peroxide [R-O-O-R, R-O-OH]	Aldehyde [R(C=O)H]
	Ketone [R(C=O)R]	Carboxilic Acid [R(C=O)OH]		Ester [R(C=O)O-R]
CHN Groups	Amine [R3N]	Hydrazine [R-NH-NH-R]	Imine [R-N=R, R-N=N-R]	Nitrile/Isonitrile [R-C ≡ N, R-N^+ ≡ C^-]
CHON Groups	NOx [R3N^+-O^-, R-N=O, R-O-N=O, R-N^+(=O)O^-, R-O-N^+(=O)O^-]			Amide [R(C=O)NR2]
CHS Groups	Thiol [R-SH]	Thioether [R-S-R]	Polysulphide [R-S-S-R, R-S-SH]	Thiocarbonyl [R(C=S)R, R(C=S)H]
CHOS Groups	SOx [R2S=O, R2S(=O)2, R-O-S(=O)-O-R, R-O-S(=O)2-O-R]			
CHX Groups (X=F, Cl, Br, I)	Halogen [R-X]			
Physical	Radical	Charges	Ionic	Solvation
	Polymer			

Security code: 38J4C7

[Type only **numerical characters**. Ignore letters and special characters.]

Search

This Search Method was already used 42 times

ThermInfo features: Searching for compounds

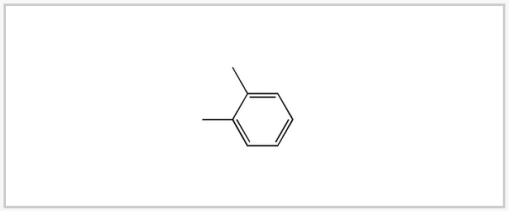
You are searching for: 1,2-Dimethylbenzene .::. Number of compounds found: 6

1. **Molecular ID:** CO01899
   Compound Name: 1,2-Dimethylbenzene
   Molecular Formula: C_8H_{10}
   CAS registry number: 95-47-6
   SMILES: Cc1ccccc1C, CC1=CC=CC=C1C
   More info: View
2. **Molecular ID:** CO01918
   Compound Name: 4-Ethyl-1,2-dimethylbenzene
   Molecular Formula: $C_{10}H_{14}$
   CAS registry number: 934-80-5
   SMILES: Cc1cc(CC)ccc1C, CC1=CC(CC)=CC=C1C
   More info: View
3. **Molecular ID:** CO01915
   Compound Name: 1-Ethyl-2,3-dimethylbenzene
   Molecular Formula: $C_{10}H_{14}$
   CAS registry number: 933-98-2
   SMILES: Cc1cccc(CC)c1C, CC1=CC=CC(CC)=C1C
   More info: View
4. **Molecular ID:** CO01999
   Compound Name: 2,3-Dimethylphenol
   Molecular Formula: $C_8H_{10}O$
   CAS registry number: 526-75-0
   SMILES: Oc1cccc(C)c1C, OC1=CC=CC(C)=C1C
   More info: View
5. **Molecular ID:** CO02008
   Compound Name: 3,4-Dimethylphenol
   Molecular Formula: $C_8H_{10}O$
   CAS registry number: 95-65-8
   SMILES: Oc1ccc(C)c(C)c1, OC1=CC=C(C)C(C)=C1
   More info: View
6. **Molecular ID:** CO02072
   Compound Name: 2,3-Dimethylbenzoic acid
   Molecular Formula: $C_9H_{10}O_2$
   CAS registry number: 603-79-2
   SMILES: O=C(O)c1cccc(C)c1C, O=C(O)C1=CC=CC(C)=C1C
   More info: View

Home | About Us | Quick Search | Advanced Search | Structural Search | Insert Data | Help | Contact Us

Monday, October 11, 2010

1,2-Dimethylbenzene



Structural Data

Molecular ID: CO01899 **Compound Name:** 1,2-Dimethylbenzene
Other Names: o-Xylene; o-Dimethylbenzene; o-Xylol; 1,2-Xylene; o-Methyltoluene; 2-Methyltoluene; UN 1307;
CASRN: 95-47-6 **Molecular Formula:** C_8H_{10}
Molecular Weight: 106.17 **Physical State:** Liquid
SMILES: Cc1ccccc1C, CC1=CC=CC=C1C
Unique SMILES: [c]1([c]([CH])[CH][CH][CH][CH]1)[CH3])[CH3], [C]1(=[C]([CH]=[CH][CH]=[CH]1)[CH3])[CH3]

Thermochemical Data

Standard Molar Enthalpy of Formation [kJ/mol]:

• Crystalline Phase:	n.a.	Error:	n.a.
• Liquid Phase:	-24.400	Error:	+/- 1.000
• Gas Phase	19.100	Error:	+/- 1.000

Standard Molar Enthalpy of Phase Change at 298.5 K [kJ/mol]:

• Solid - Liquid	n.a.	Error:	n.a.
• Liquid - Gas	43.400	Error:	+/- 0.000
• Solid - Gas	n.a.	Error:	n.a.

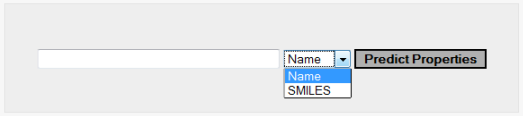
Observations: n.a.

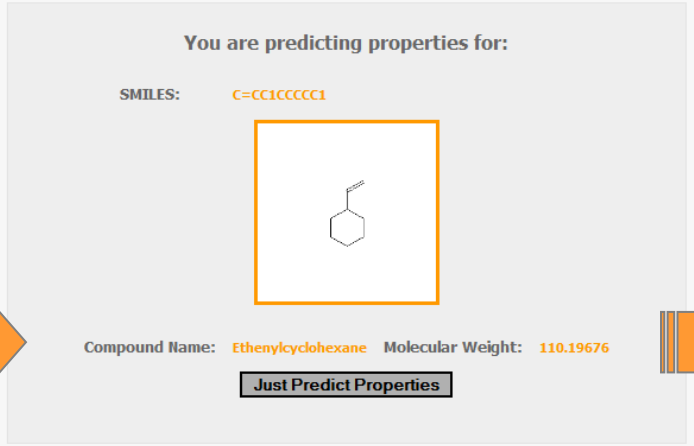
References

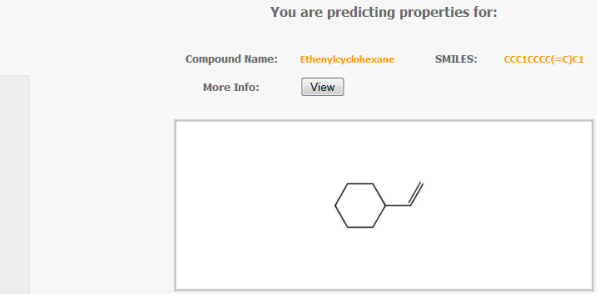
1. J. B. Pedley
   Thermochemical Data and Structures of Organic Compounds, Vol. 1, 1994, 571
2. J. B. Pedley, R. D. Naylor, S. P. Kirby
   Thermochemical Data of Organic Compounds, 2nd ed., 1986, 792

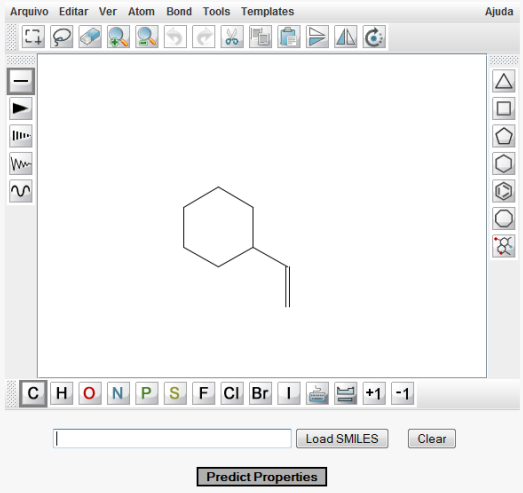
Print this Page | Save Page as PDF

ThermInfo features: Predicting properties



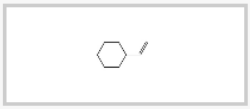




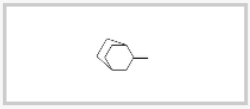


.:: Number of compounds found on the database: 2 ::.

* Experimental and Estimated Properties for similar compounds found on the database: *



Molecular ID: CO01625
Compound Name: Ethenylcyclohexane
Molecular Formula: C_8H_{14}
Molecular Weight: 110.20
SMILES: C=CC1CCCCC1
Properties Prediction: View



Molecular ID: CO02671
Compound Name: 2-Methylene-bicyclo[2.2.2]octane
Molecular Formula: C_9H_{14}
Molecular Weight: 122.21
SMILES: C=C1C(CC2)CCC2C1
Properties Prediction: View

Experimental Standard Molar Enthalpy of Formation at 298.15 K for C=CC1CCCCC1 [kJ/mol]:

Gas-phase:	Liquid-phase:
-48.90	-88.70

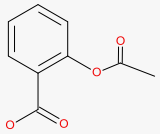
Estimated Standard Molar Enthalpy of Formation at 298.15 K using ELBA for CCC1CCCC(=C)C1 [kJ/mol]:

Gas-phase:	Liquid-phase:
-68.77	-114.3

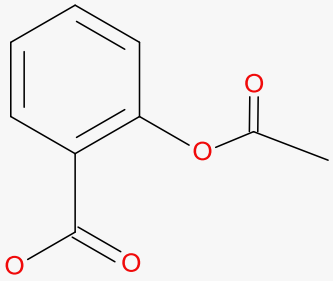
Set of ELBA parameters used to predict properties of CCC1CCCC(=C)C1

ELBA parameter	Use Frequency	Short Description [More Info]
C2C2	2	bond between two carbon atoms bonded to two carbon atoms [C-C bond for Alkanes]
C2C3	3	bond between a carbon atom bonded to two carbon atoms and a carbon atom bonded to three carbon atoms [C-C bond for Alkanes]
C1H	3	bond between a hydrogen atom and a carbon that is bonded to one carbon atom [C-H bond for Alkanes]
C3H	1	bond between a hydrogen atom and a carbon that is bonded to three carbon atoms [C-H bond for Alkanes]
C2D3	2	bond between a carbon atom bonded to two carbon atoms and a carbon atom involved in a double bond and bonded to three carbon atoms [C-C bond for Alkenes]
ZS6C2	4	secondary carbon atom in a six-carbon ring [strain parameter for cycloalkanes]
D1D3	1	double bond between a carbon atom bonded to one carbon atom and a carbon atom bonded to three carbon atoms [C-C bond for Alkenes]
ZS6D3	1	double bond connected directly to a six-carbon ring [strain parameter for cycloalkanes]
ZS6C3	1	tertiary carbon atom in a six-carbon ring [strain parameter for cycloalkanes]
C1C2	1	bond between a carbon atom bonded to one carbon atom and a carbon atom bonded to two carbon atoms [C-C bond for Alkanes]
C2H	10	bond between a hydrogen atom and a carbon that is bonded to two carbon atoms [C-H bond for Alkanes]
D1H	2	bond between a hydrogen atom and a carbon that is only involved in a double bond [C-H bond for Alkenes]
Z6	4	six-atoms ring having: one double bond in the ring, in the presence or absence of one substituent (Z6 is multiplied by two). one double bond in the ring, in the presence of two substituents (Z6 is multiplied by four). two conjugated double bonds in the ring, in the absence of one substituents (Z6 is multiplied by eight). two conjugated double bonds in the ring, in the presence of one or two substituents (Z6 is multiplied by five)[conformational parameter for cycloalkenes]

Prediction of thermochemical properties

Property = *f*()

Molecular Structure → **Molecular Representation** → **Machine Learning** → **Properties**



- SMILES O=C(Oc1ccccc1C(=O)O)C
- Molecular Fingerprint 1000010000000110000010010 0000001
- Similarity of Molecular Fingerprints
- Molecular Descriptors: 0D, 2D, 3D
- Connection table
- Adjacency matrix
- ...

- Neural Networks
- Support vector machines
- PCA
- ...

$\Delta_f H°_m$ = -815.6 kJ/mol

…

"The future of chemistry depends on the automated analysis of chemical knowledge, combining disparate data sources in a single resource which can be analyzed using computational techniques to assess and build on these data."

Townsend et al. (2004)

http://www.therminfo.com

Hypertrophic cardiomyopathy: how can the Semantic Web and data mining help?

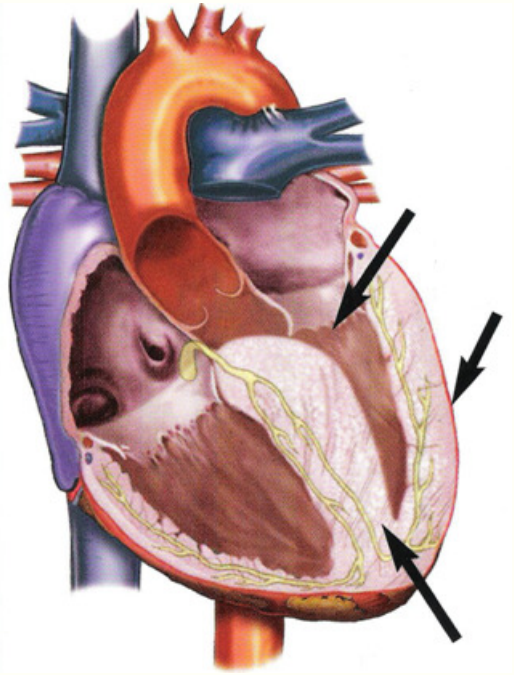
Cátia M. Machado

May 16th, 2011

Hypertrophic cardiomyopathy (HCM)

- Hereditary cardiovascular disease

- The most frequent cause of sudden cardiac death among apparently healthy young people and athletes



Hypertrophic cardiomyopathy (HCM)

Clinical evaluation

- Clinical history: e.g. cardiac auscultation, family history
- Exams: e.g. echocardiogram, electrocardiogram

Genetic analysis

- 20 Genes
- > 640 Mutations

Diagnosis & Prognosis

- Difficult diagnosis before onset of severe symptoms
- Differential diagnosis
- Unknown prognosis

- 1 Mutation = positive diagnosis
- No mutation ≠ negative diagnosis
- Different symptoms with the same mutation

The Semantic Web and its technologies

Semantic Web (SW)

Web of documents ⟹ Web of data

- Meaningful data elements
- Machine-processable
- Technologies as standards: everyone "should" use them
- Framework for data integration

SW technologies

1. RDF: Language for data representation and interchange on the Web

(Subject) (Predicate) (Object)

The Semantic Web and its technologies

2. RDFS & OWL: RDF schema & Web Ontology Language

Advantages:

- Better description of the domain
- Inference

How:

E.g. Class = Person

Subclass = Adult

Jeff type Adult

Advantages for the HCM study

- Explicit representation of the disease and its diagnosis
- Easier integration with data already in the (Semantic) Web

Data mining

Objectives

- Identification of correlations between phenotype and genotype

Normally done with statistical algorithms – good results with a small number of variables

State of the art

- Support vector machines: breast cancer susceptibility
- Neural networks: gastric cancer recurrence

Advantages for the HCM study

- Ability to deal with a high number of variables
- Predictive power vs. expressiveness (e.g. decision trees)

Analysis of Protein Families
GO Annotations

- Protein functional annotation can vary in:

  - coverage (regarding a family or set of proteins);
  - type (manual, automatic);
  - quality;
  - etc.

Analysis of Protein Families
GO Annotations

- GRYFUN (http://xldb.fc.ul.pt/biotools/gryfun/)

  [**GR**aph anal**Y**zer of **FUN**ctional annotation]

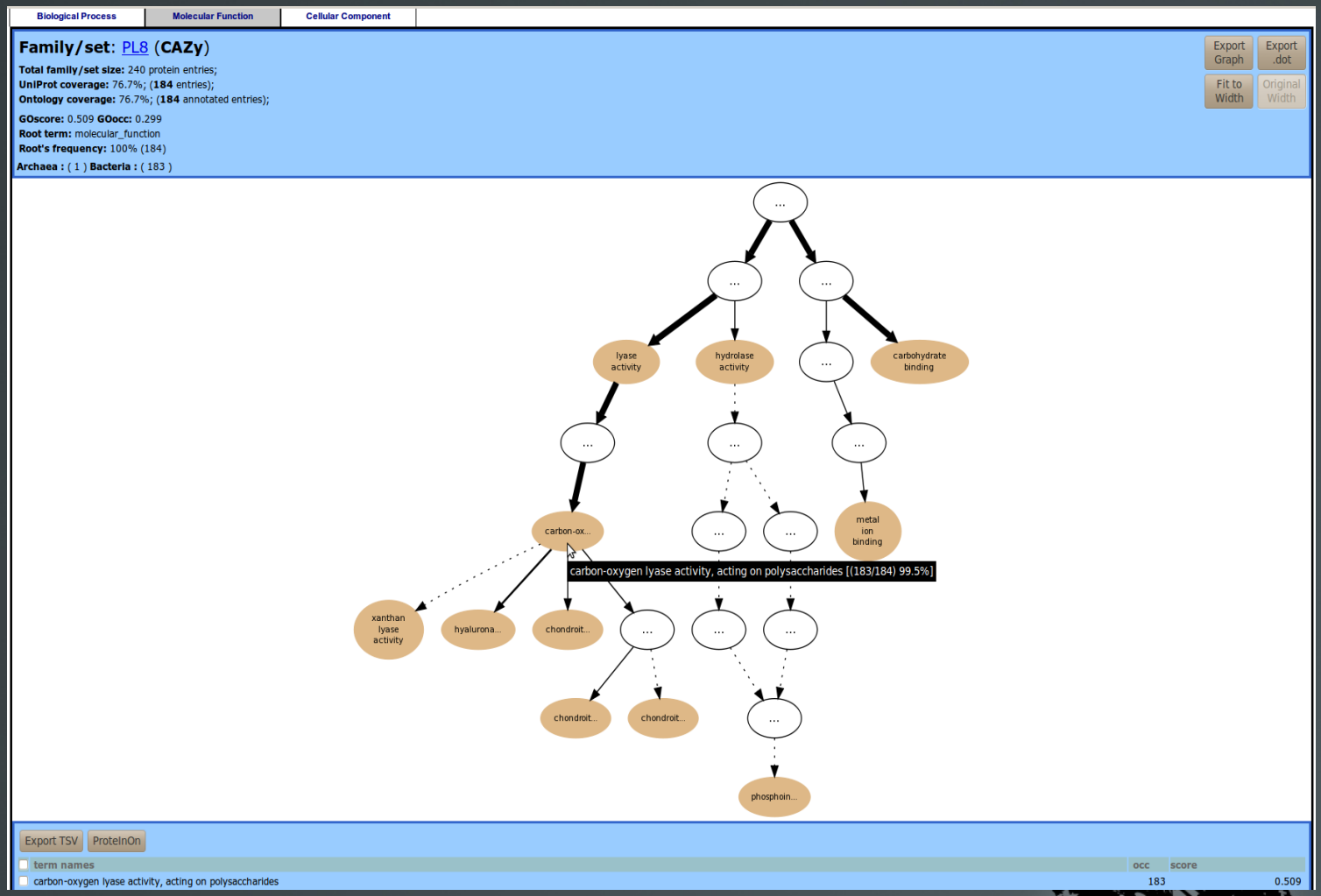
- is a webtool that allows for the visualization, filtering and subsequent analysis of Gene Ontology (GO) annotation profiles

Analysis of Protein Families
GO Annotations

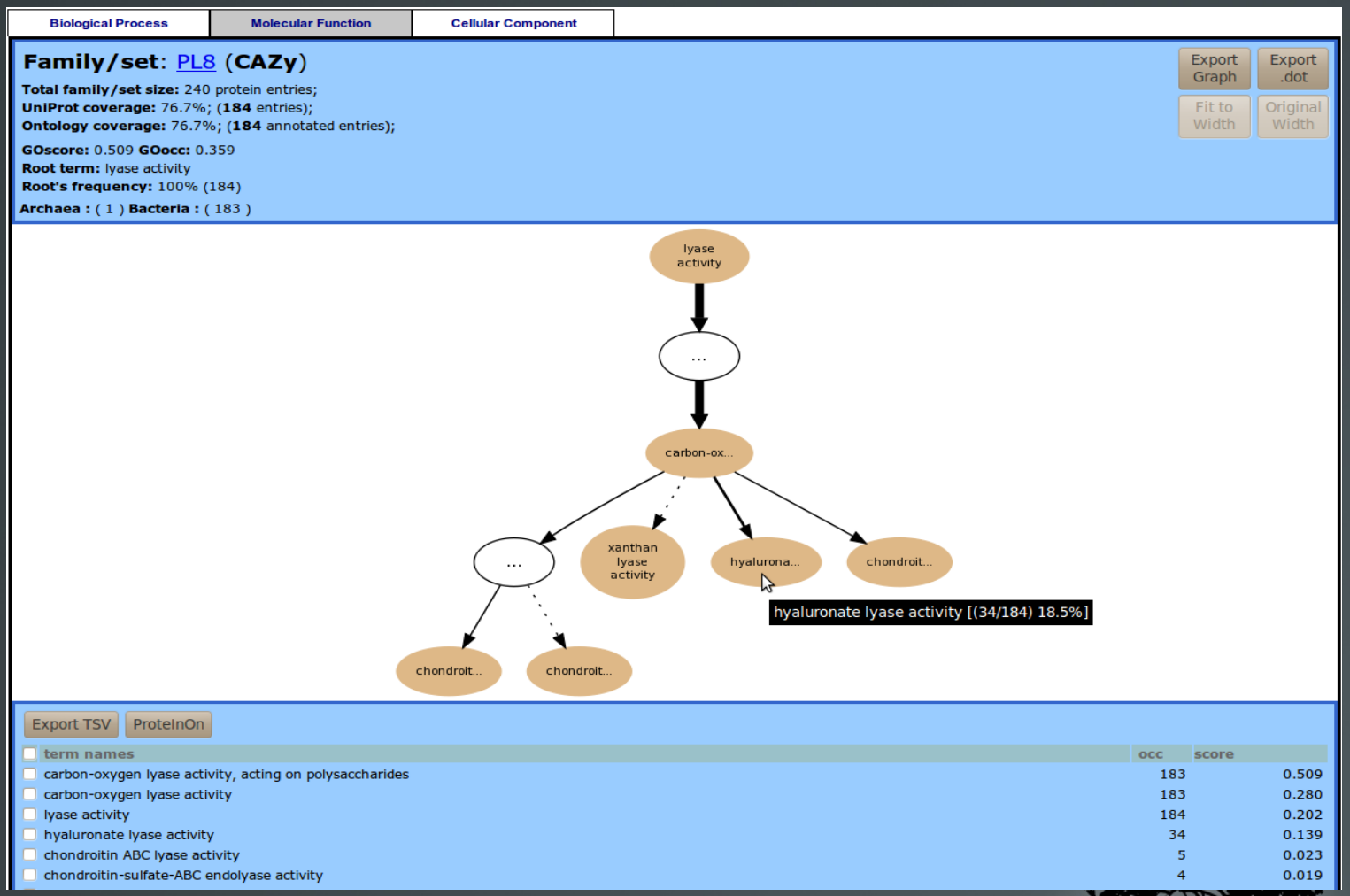
- Gene Ontology (**GO**) provides

  - a structured and controlled vocabulary,
  - with 3 orthogonal ontologies (**biological process**, **molecular function** and **cellular component**),
  - which facilitates functional analyses and development of metrics

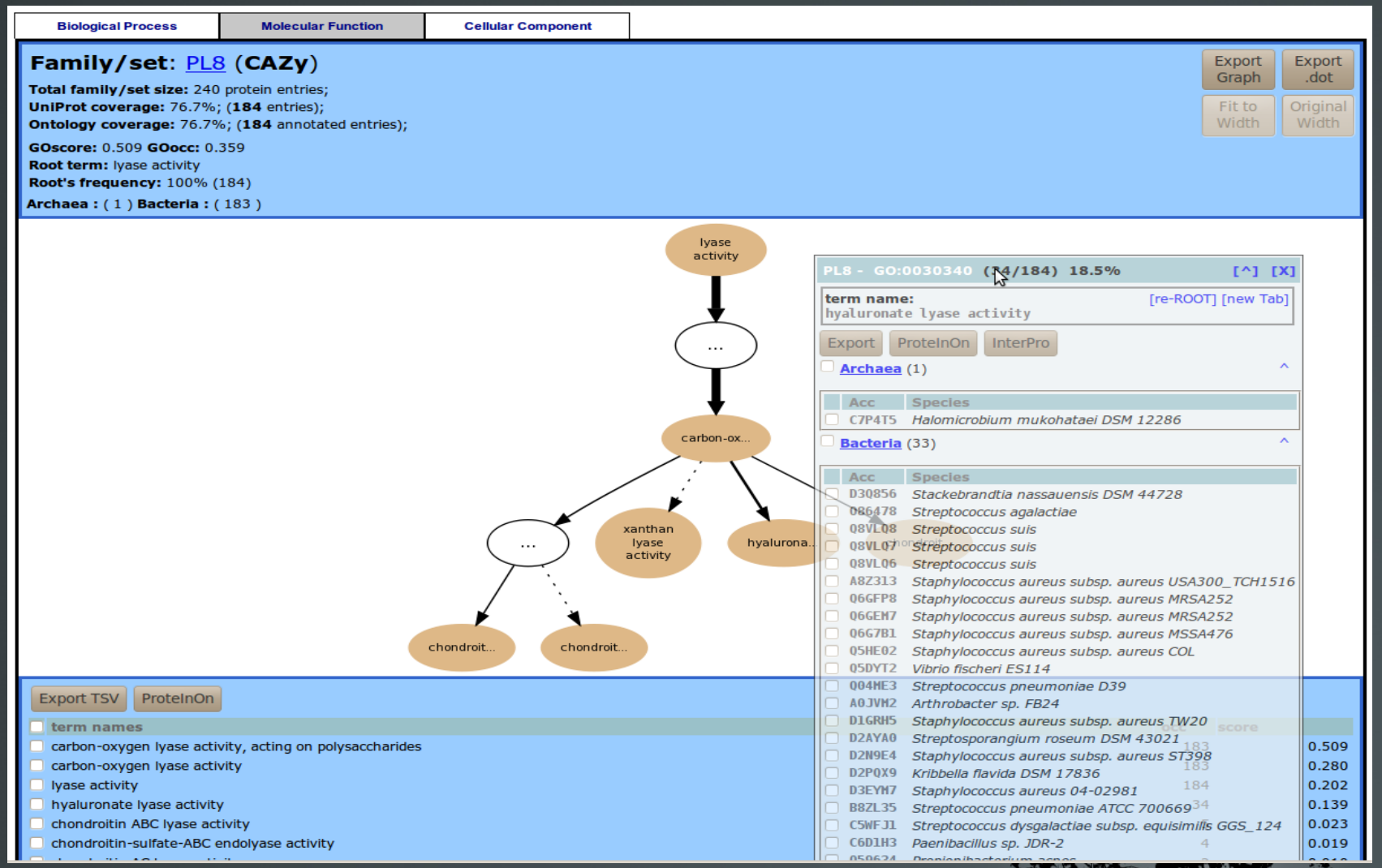
Analysis of Protein Families
GO Annotations



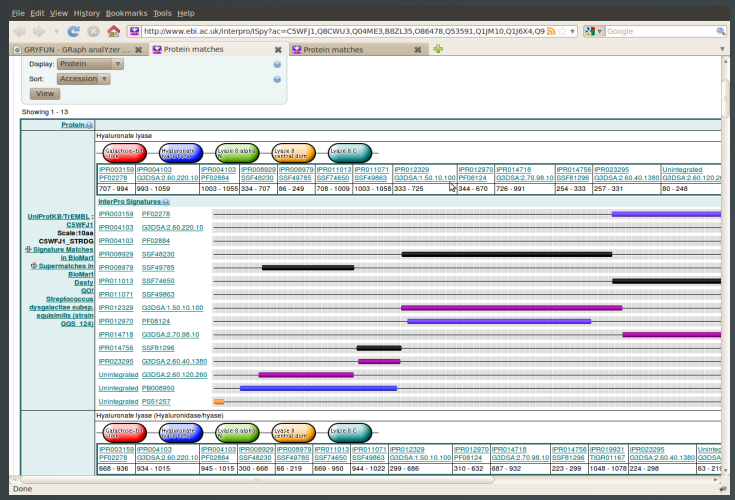
Analysis of Protein Families
GO Annotations

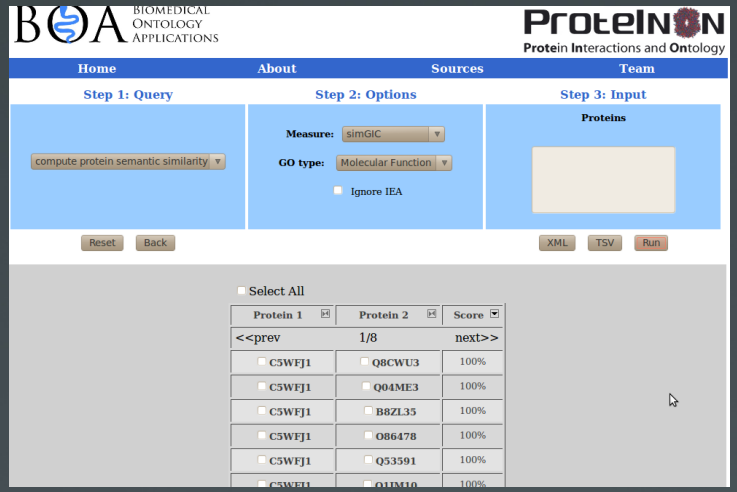


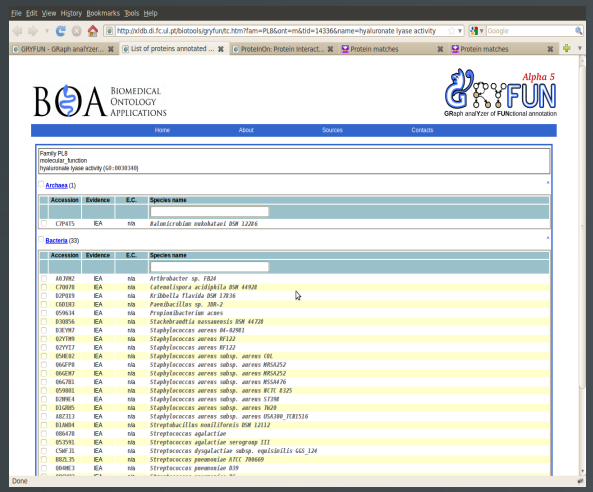
Analysis of Protein Families
GO Annotations

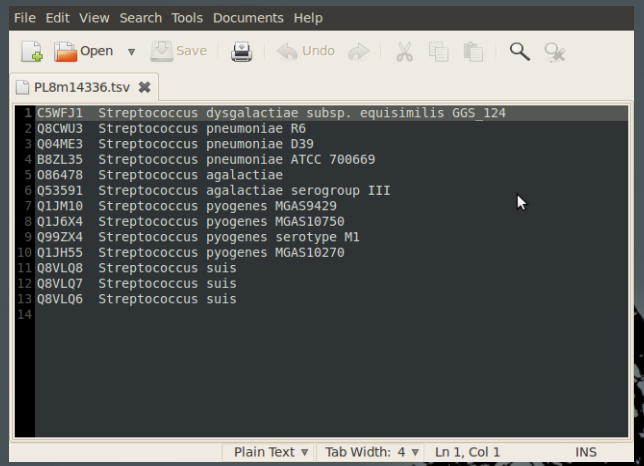


Analysis of Protein Families
GO Annotations







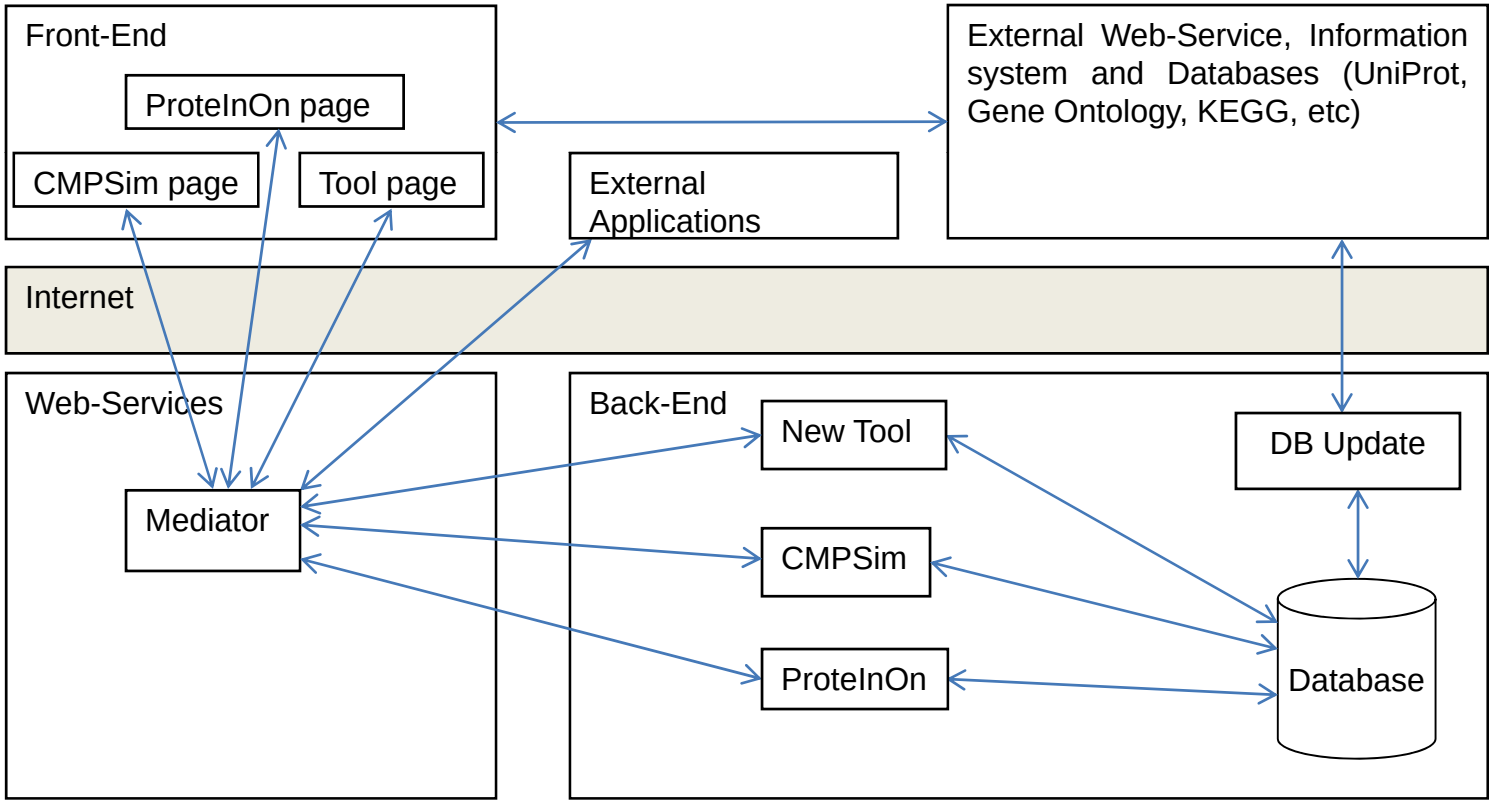


Biomedical Ontology Applications Framework

16/05/11

Bruno Tavares

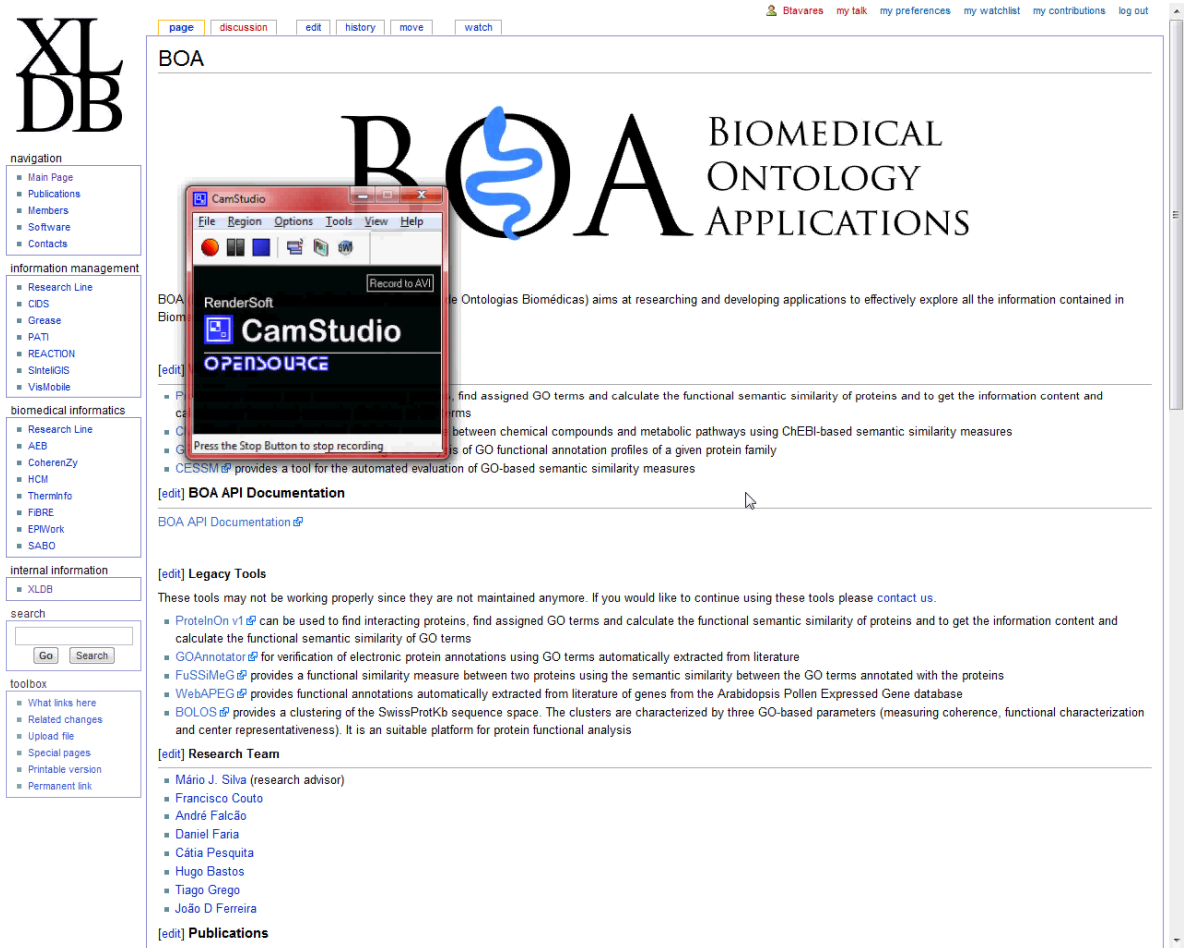
Biomedical Ontology Applications Architecture



WebServices

- RESTfull (Representational State Transfer) architecture
- RPC (remote procedure call) architecture

Demonstration Video



Aprendizagem automática para previsão de permeabilidade de compostos na membrana hemato-encefálica

Machine learning algorithms to predict blood-brain barrier permeability of drug molecules

Inês Filipa Martins

XLDB-LaSIGE
Departamento Informática
Faculdade de Ciências da Universidade de Lisboa

Barreira hemato-encefálica

Membrana presente no encéfalo, constituída por células endoteliais justapostas.

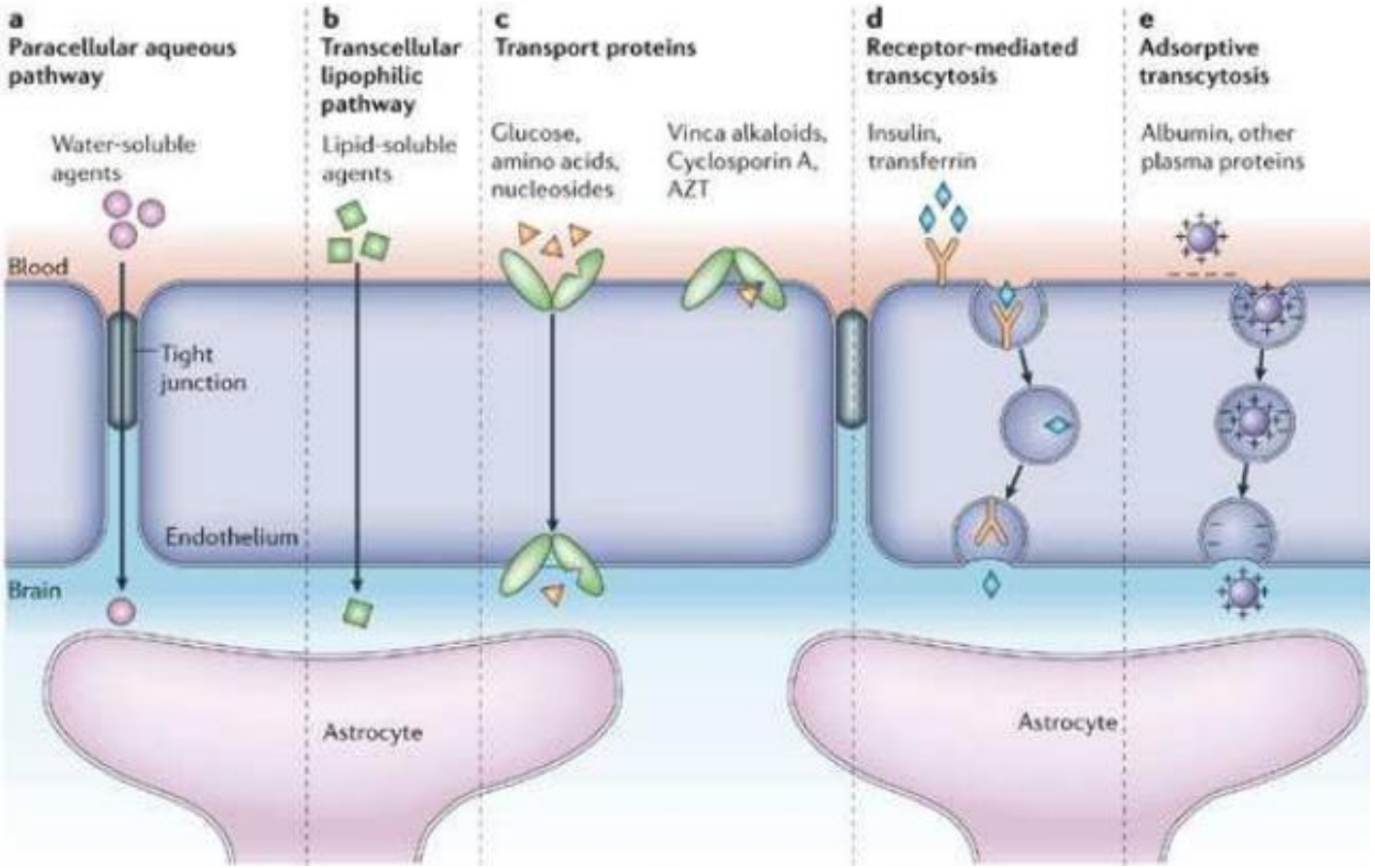
Previne a passagem de substâncias químicas da corrente sanguínea para o cérebro.

Maior restrição quando comparamos com células endoteliais presentes no resto do organismo.

Permeabilidade selectiva:

Substâncias apolares
Lipofílicas
Tamanho molecular reduzido
Elevado coeficiente de partição octanol/água (logP)

Transporte através da BHE

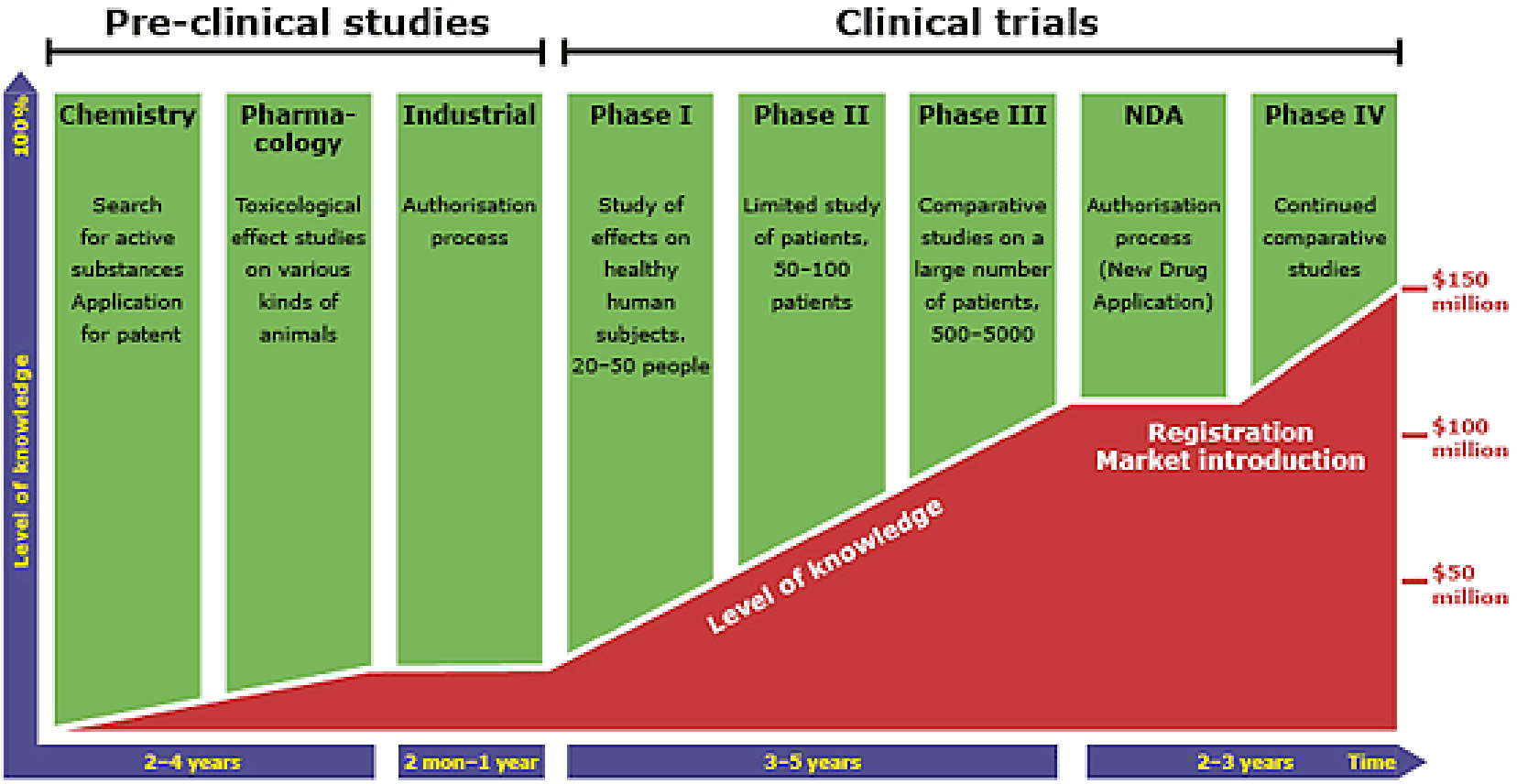


Difusão
Transportadores proteicos
Endocitose mediada ou não por receptores.

Fases desenvolvimento de um fármaco

Fase pré-clínica

Pesquisa de um fármaco promissor para uma determinada patologia (nesta fase são testadas **milhares de substâncias**, sendo que poucas são seleccionadas para estudos posteriores).



Objectivo do trabalho

Testar metodologias para determinar com melhor precisão se um determinado fármaco pode passar a membrana hemato-encefálica

. Similariedade estrutural de moléculas

. Utilização de informação estrutural das moléculas

. Não são consideradas as propriedades da membrana

Aprendizagem automática

Desenvolvimento de métodos computacionais que permitem aos computadores "aprender" com a experiência.

Extrair informação útil de grandes volumes de dados.

Descobrir padrões de interesse, comportamentos, tendências.

Abordagem

Base de dados de compostos - 628 moléculas

Fingerprints

Codificam moléculas numa string de bits (1 e 0) que representam a presença ou ausência de uma sub-estrutura da molécula.

A comparação de fingerprints permite determinar a similaridade entre 2 moléculas, medida pelo **coeficiente de Tanimoto**

. proporção de *bits on* (1) que são partilhados entre as moléculas

Resultado: matriz de medidas de similaridade 628x628

Abordagem - continuação

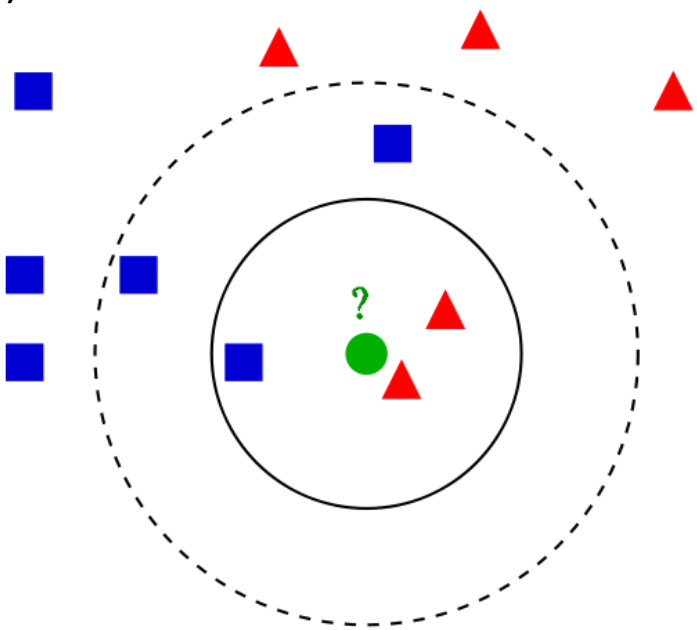
Escalamento multidimensional linear (análise em coordenadas principais)

. matriz de medidas de similaridade
. Encontra a melhor representação geométrica dos dados no espaço
. Transforma matriz inicial num vector de coordenadas para cada composto

K-NN (k-Nearest Neighbour)

. Variação do número de vizinhos (1, 2, 4 e 8)

. Classificação das moléculas de acordo com a classificação das moléculas vizinhas



Moléculas são divididas em:
Treino
Teste

Abordagem - resultado

Accuracy - proporção de resultados classificados correctamente.

Phi – coeficiente de correlação entre o observado e o previsto em classificações binárias (passa - 1, não passa - 0).

Valor entre [-1, 1]

previsão inversa previsão perfeita

Na literatura…

Accuracy 73% → Valor a ultrapassar!

Utilizando outros métodos e propriedades

Annotating protein functions with n-peptides using data mining techniques

A new way to say what an unknown protein might do just by cutting and counting it

Manuela Cardoso
Mestrado em Bioinformática
e Biologia Computacional

XLDB-LaSIGE: DI-FCUL
2010/11

Intro

- Frequência de alguns VSPS (*very small protein sequences*, de 3 a 5 aa) desviam-se do esperado.
  - Há alguma razão para tal?
  - Como pode ser explorado?

- A atribuição de função a proteínas é uma tarefa com grandes custos a nível temporal e financeiro.
  - Necessidade de novas formas rápidas, fáceis e credíveis de as anotar ou confirmar anotações existentes.

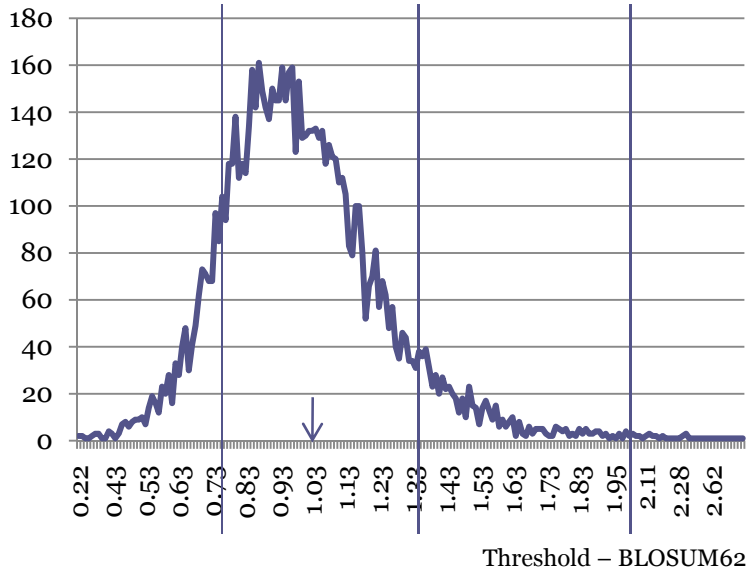
Variação das frequências de tripéptidos

Bacteria

$$\frac{\text{freq (obs)}}{\text{freq (esp)}}$$

	#	freq(#)
> 2,000	53	0.006625
> 1,330	709	0.088625
< 0,750	947	0.118375
normais	6291	0.786375
	8000	1

10 menores			10 maiores		
	CKW	0.223382		HHP	2.62403
	KCM	0.228715		WQC	2.729661
	YMW	0.297737		WQQ	2.970405
	IMW	0.297982		HCW	3.034376
	CMC	0.335434		WHW	3.135888
	CMR	0.36208		CGC	3.158789
	HKC	0.375328		HHH	3.335922
	MWK	0.378816		CPC	3.592887
	KCA	0.38186		HCH	3.753144
	CMW	0.388049		CCW	4.226074



Threshold – BLOSUM62

Uniprot Taxonomy

- Utilização de tripéptidos
  - Suficiente informação
  - Não em demasia, arriscando ter combinações muito raras

Combinações = 2^n

- Função molecular

Objectivo

Desenvolver uma ferramenta com a qual se poderá atribuir funções moleculares a uma dada proteína, através do seu conteúdo em tripéptidos

“Ferramentas”

- Bases de dados (BD):
  - Gene Ontology (GO)
    - Uniprot (web)
    - Mygo (servidor)
- Programação
  - Python 2.7
  - mySQL
  - Microsoft Excel

Grupo teste (1)

- BD mygo v.2010
  - Dados suficientemente “antigos” a fim de ver se os métodos conseguiram prever inserção ou remoção de funções

- 30 proteínas
  - Evidência EXP + IDA
    - Anotações mais credíveis para puder comparar
  - Aleatória

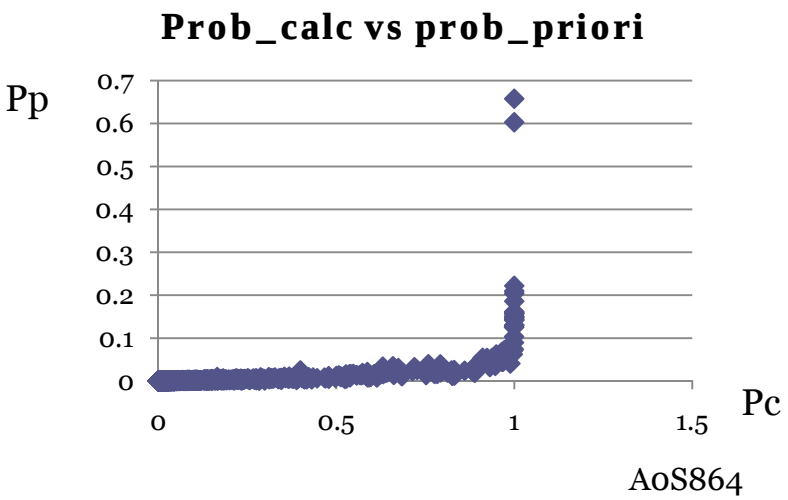
Mecanismo geral

- Dada uma sequência proteica:
  - Programa em Python “corta-a” em tripéptidos e conta as ocorrências
  - Compara com os dados da BD
    - Quais as funções têm as proteínas com esses tripéptidos, valores obs vs esp, etc
  - Calcula e estima as probabilidades associado a cada função possível dessa proteína
  - Retorna lista de funções moleculares, ordenada decrescentemente por probabilidade

Estatística

- Baseado no teorema de Bayes simples

$$P(term|trip) = \frac{P(trip|term)P(term)}{P(trip)}$$



Probabilidade calculada segue tendencialmente a probabilidade à priori

↓

Pondera demasiado alto as funções mais comuns

↓

Problema!

- “if”
  - Se rácio obs/esp ≥1,
    - Rácio – 1
  - Senão,
    - 1-1/rácio
- log”
  - Log(rácio obs/esp)

- Em ambos:
  - Somou-se valores pertencentes à mesma função (devido aos diferentes trip puder “possuir” iguais funções)
  - Valor central (quando obs ≈esp) = 0
  - Se > 0, teste diz ser provável a proteína ter essa função
  - Se < 0, teste diz ter prob negativa

- Tabela resultados com “if” e “log”
  - Só inclui funções que a prot A0S864 tem

acc	go_id	prob_priori	sum(if)	sum(log)	ord_sum(if)	ord_sum(log)	go_name
A0S864	3712	0.0446535	-3.467306883	-1.10726	1034	969	signal transducer activity
A0S864	3903	0.000618355	35.52870137	7.962272	438	196	receptor signaling protein activity
A0S864	14428	9.46E-05	303.0668428	21.10216	5	5	receptor regulator activity
A0S864	14430	8.16E-05	669.8192708	28.7977	3	3	receptor inhibitor activity
A0S864	14431	7.48E-05	697.5094116	31.5339	2	2	acetylcholine receptor regulator activity
A0S864	14433	6.99E-05	1063.690073	36.98107	1	1	acetylcholine receptor inhibitor activity

De modo geral, funções raras estão ficam mais no topo

“Log” tem melhor desempenho

Grupo teste (2)

- 100 proteínas
  - Sem evidências EXP nem IDA
    - Testar se os resultados anteriores se devem ao facto dessas proteínas serem limitadas e afectarem outras por vias de anotação electrónica.
      - Eg: classificar erradamente proteínas estruturalmente semelhantes como tendo a mesma função
  - Aleatório do mygo v.2010

A correr neste momento...

A seguir...

- Ferramenta online disponível para uso da comunidade científica